AFIT/GIR/LAL/97D-2

DEVELOPING A CORPUS SPECIFIC STOPLIST USING QUANTITATIVE COMPARISON

THESIS

Craig N. Berg
Captain, USAF

AFIT/GIR/LAL/97D-2

Approved for public release; distribution unlimited

The views expressed in this thesis are those of the author
and do not reflect the official policy or position of the
Department of Defense or the U.S. Government.

AFIT/GIR/LAL/97D-2

DEVELOPING A CORPUS SPECIFIC STOPLIST
USING QUANTITATIVE COMPARISON

THESIS

Presented to the Faculty of the Graduate School of Logistics

and Acquisition Management of the Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Information Resource Management

Craig N. Berg, B.S.
Captain, USAF

December 1997

Acknowledgments

I would like to take this time to acknowledge all of the support and understanding my loving family provided to me during the past 19 months. Without their help this would have been a much more difficult effort. To my wife Susan, who many times gave me the encouragement I needed to get going when I really didn't want to, I will never be able to truly express how much I appreciate or love you.

I would also like to thank my advisor, Dr. Kim Campbell, and my reader, Prof Dan Ferens; their many suggestions and comments led directly to the product you now see.

A special thanks go to Judy Berger and Mark Steinberg from the Defense Technical Information Center, Office of Data Administration. Their assistance and hours of effort helped me obtain the data I needed to perform my analysis.

Craig N. Berg

Table of Contents

Page

Page

Page

List of Figures

Figure Page

List of Tables

Table Page

AFIT/GIR/LAL/97D-2

Abstract

We have become overwhelmed with electronic information and it seems our situation is not going to improve. When computers first became thought of as instruments to assist us and make our lives easier we thought the future would be a manageable one. We envisioned a day when documents, no matter when they were produced, would be as close as a click of the mouse and the typing of a few words. Locating information of interest was not going to take all day. What we have found is technology changes faster than we can keep up with it.

This thesis will look at how we can provide faster access to the information we are looking for. Previous research in the area of document/information retrieval has mainly focused on the automated creation of abstracts and indexes. But today's requirements are more closely related to searching for information through the use of queries. At the heart of the query process is the removal of search terms with little or no significance to the search being performed. More often than not stoplists are constructed from the most commonly occurring words in the English language. This approach may be fine for systems which handle information from very broad categories. We will examine how to build a stoplist for a specific area of interest, such as the Air Force, based on a specific body of linguistic data, or corpus. The stoplist developed from the Air Force corpus was also tested to see if it provides a quicker response than the generic stoplist when used with the United States Air Force's Official homepage, http://www.af.mil.

DEVELOPING A CORPUS SPECIFIC STOPLIST USING QUANTITATIVE COMPARISON ANALYSIS

I. Introduction

Chapter Overview

The Air Force, and the world as a whole, is being inundated with information. Today most of this information comes in electronic form and reduces the burden somewhat, or does it? We have made leaps and bounds in our pursuit of technology. Gone are the manual typewriters and mimeograph machines of the last decade, and we look at this as an advancement. Ask most people if they can find the information they are looking for, and they may tell you they resort to the same methods they used to use with their "old-fashioned" office equipment.

In short, we have created the machinery to produce, distribute and reproduce more information than ever before, but have not improved the methods for processing this information. The older methods, designed to handle paper-based information, will not work with electronic information, and we have not created a new way to deal with the massive quantities of information. Since technology is moving at such a fast pace we often find ourselves using what worked in the past, and seems to work right now.

We are storing more and more information in computer databases every day. We store information for later retrieval, expecting it to be there when, and if we need it, and we hope we can obtain it in an efficient and timely manner. When we only had to go to a filing cabinet to get what we needed we never thought about how long it would take even when the cabinet was full. Even after moving to a desktop computer and Windows based applications most people could tell you where they put each piece of information they created. We now face a massive amount of information obtainable through the Internet. We have linked millions of computers to create one large family, and opened up access to unknown number of sources for everyone to use. Yes, it is easy to find a method to locate what you are trying to find. Web search engines such as Yahoo!, Excite, Hot Bot, and Metacrawler have provided an easy means of finding just what you need. Or have they? It is quite conceivable to wait at your computer for several minutes while your pet search engine combs its massive catalog of information looking for what you have requested. Even more extreme is the thought of obtaining information from a large data warehouse such as LEXIS-NEXIS in Dayton OH. In its data warehouse LEXIS-NEXIS maintains more than 1.2 billion documents, consuming 2.5 terabytes (2.5 trillion characters) of storage, of source information for its subscribers to access (LEXIS, 1997). Providing timely access to this information is a main concern to LEXIS-NEXIS, as it is to everyone who stores information with the hope of one day retrieving it.

The act or process of storing and retrieving information is called information retrieval. An Information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request (Van Rijsbergen, 1979:1). Information retrieval can be as simple as opening a file on a single floppy disk for a home user. Information retrieval can also be a process spanning multiple systems across geographic areas if you are searching for information maintained in an airline's reservation system.

In the late 1950's it was envisioned that computers would be able to take entire documents and automatically create an abstract or an index of the document (Luhn, 1958:159). A considerable amount of effort was put into producing computer programs that would catalog and index all the information we were producing. The intent was to use computers to replace humans who normally performed the functions of cataloging and indexing. The idea of indexing every document has been abandoned due to the volume of information we are creating. Today we store entire documents, not abstracts of them, and users must find them through whatever method they choose. We now rely very heavily on search engines as a means of accessing information located in places other than our computer. This applies to information accessed on or through the Internet as well as information held in large corporate databases. As the amount of information increases it will become more difficult to find what we are looking for; as this happens the search engine and search techniques will become more important to everyone.

Search Engines and Searching

To have a reasonable discussion of searching and search techniques it is necessary to establish a basic groundwork. This research effort is going to concentrate on a very small area of search techniques, specifically web page searching. In particular, we will examine the operation of the United States Air Force homepage.

Document Storage

Each document accessible on a website is cataloged to allow easy access and timely retrieval. This process is carried out on a regular basis by the webpage's search software, and does not require intervention by a system administrator. A system administrator can make changes to the efficiency of the process and adjust certain parameters of the process if needed. Figure one outlines the process described above. Each articles is read in by the catalog process, and

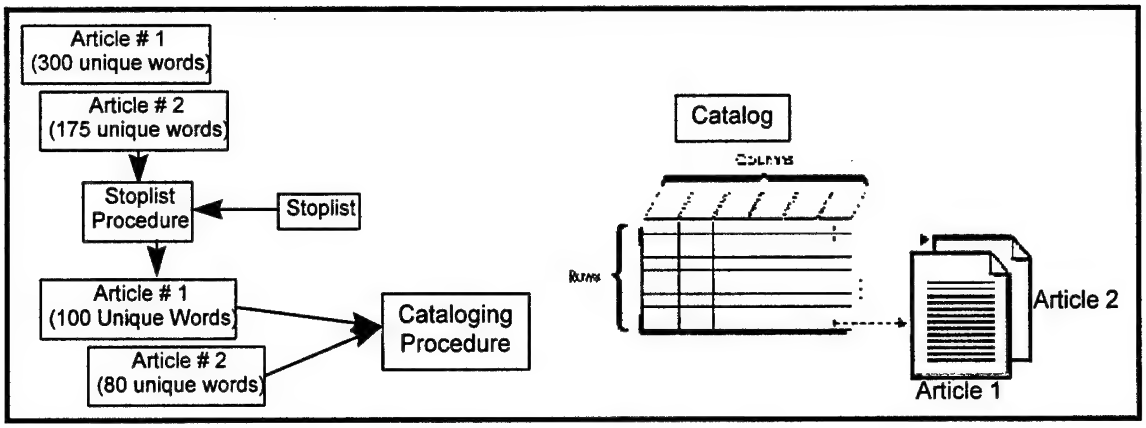


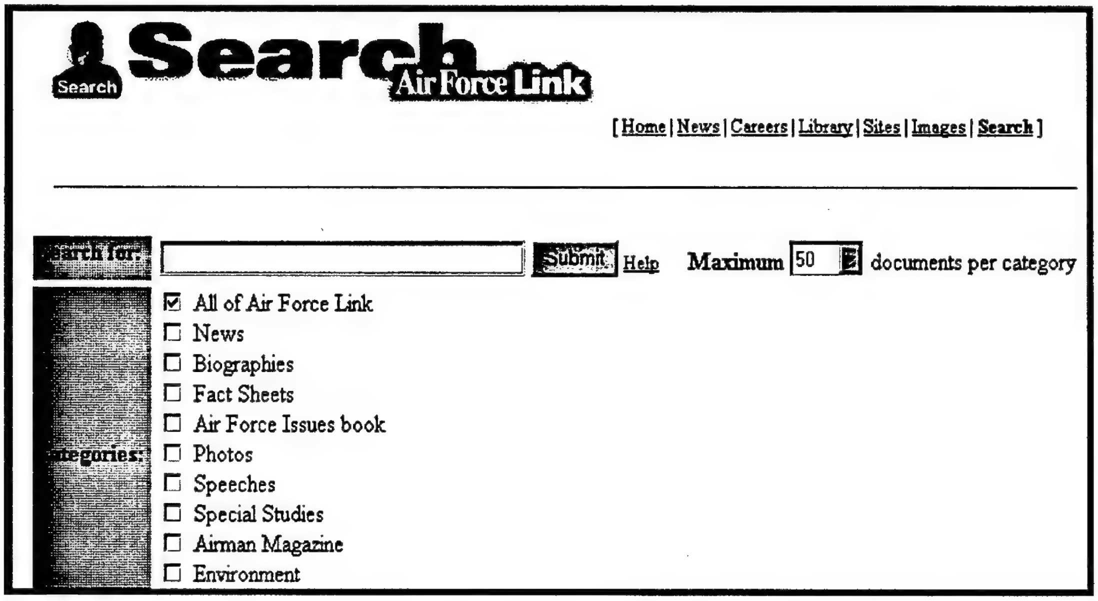
Figure1. Webpage information cataloging process

stoplisted. Stoplisting removes all words contained in a list called a stoplist. A stoplist, or negative dictionary, is a list of words that considered to have little significance in terms of their searching power (Fox, 1989:19). A search engine has a default stoplist, normally containing the most commonly occurring words in the English language. The stoplist can be altered by a system administrator to contain words they feel are important. The stoplist can significantly reduce the amount of information to be cataloged on a webpage. In this example the number of unique terms in Article # 1 is reduced by 41.67 percent and the number of unique terms in Article # 2 is reduced by 20 percent. The document is stoplisted, the important information is located, and the location of the information is recorded in the catalog. A search engine may catalog either individual words or only certain items such as the title, author, or the first ten lines. Regardless of the amount of information cataloged, the a link is maintained between the catalog entry and the location of the source document.

The catalog is now ready to be used by the query process to locate all pieces of information that might contain the information being requested.

Query Operations

The basic function of a query, whether it is done on the world wide web or on a local database, is to determine the location of any piece of information meeting the condition(s) specified by the requester. As an example, figure 2 shows the query interface for the United States Air Force's homepage. The input screen allows the user to specify what they would like to search for, what areas they would like to look in, and to place a limit on the number of documents to be retrieved in each category. Other search engine interfaces allow the user to specify if the terms in the query are to be treated as a phrase, if all terms must be found in the document, or if just some of the query terms can be contained in the document. Viewed as a block diagram process, figure three shows how the


Search
Search
Air Force Link
[Home | News | Careers | Library | Sites | Images | Search]
earch for:
Submit
Help
tegories:

Maximum 50 documents per category

☑ All of Air Force Link
□ News
□ Biographies
□ Fact Sheets
□ Air Force Issues book
□ Photos
□ Speeches
□ Special Studies
□ Airman Magazine
□ Environment

Figure 2. Query input screen from US Air Force homepage http://www.af.mil/search.htm

search query is processed. The catalog is searched for the specific information

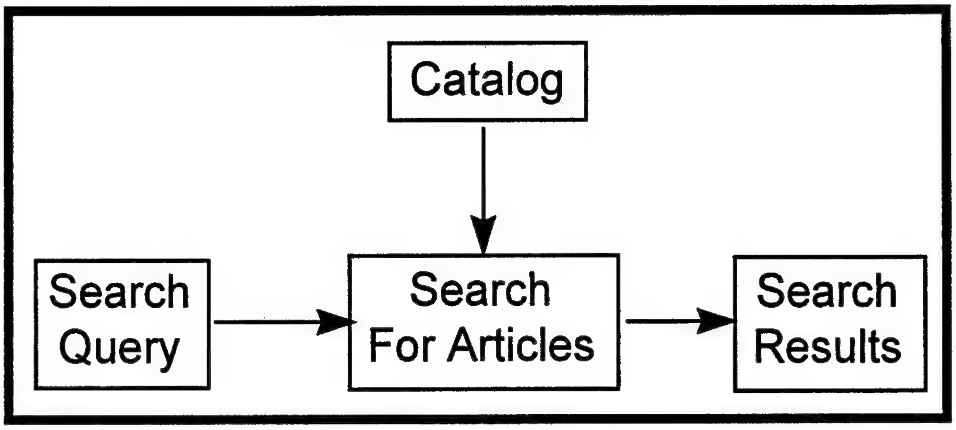


Figure 3. Query Process

contained in the search query, and the corresponding search results, the location and description in some cases, are provided to the user.

Problems with Querying

Cataloging information and the query process seem to work very effectively and provide information promptly, at least today. Depending on what you are looking for, and how you search for it, a query can sometimes take a bit longer than we might expect. One reason for a delay might be the size of the catalog being searched. When searching a homepage, such as the US Air Force homepage, the size of the catalog is directly proportional to the number of articles accessible and the number of words within those documents. It is rather unreasonable to limit the number of words in a piece of information, and it is just as unreasonable to limit the amount of information a homepage may contain. Another option is to control the number of words stored within the catalog of the homepage. By reducing the number of words stored in the catalog the size of the catalog will be reduced and the time required to search through the catalog, matching the query terms with words in the catalog, will be reduced. Limiting the number of words in the catalog can best be done by using a stoplist.

Stoplists

Our examples above show a stoplist is already used to remove words believed to be of little importance in creating an index. In fact, stoplists have generally been composed of the most commonly occurring words in general use (Fox, 1989:19). Hompages and large databases usually consist of information from a more narrow field of interest, and will probably use a different basic vocabulary than seen in general conversation or writing. It is conceivable a stoplist based on the frequency of word usage in a specific subject area would differ from one based on general word usage. If such a subject area specific stoplist could be created it reasonable to assume the size of the information catalogs for a homepage could be decreased, and the speed of query operations.

Problem Statement

A well-defined method for creating a subject area, or corpus based, stoplist does not exist. Because technology in the area of computers has progressed quicker than research in the area of computer applications we have forced users to make-do with a commonly accepted list of "most frequently used words." The list of most frequently used words is based upon a generic corpus of written English. A corpus is "a collection of naturally occurring language text, chosen to characterize a state or variety of a language" (Sinclair, 1991).

Research Objectives

The following research objectives must be met to solve the specific problem of interest:

- Locate and summarize the documented methods for creating a stoplist from a corpus.
- Develop and test a simple (2-3 step) process for creating a stoplist tailored to a subject area.
- Determine the effectiveness of a corpus specific stoplist through statistical comparison.

Research Hypotheses

A. The words contained in a stoplist generated from an Air Force corpus will be significantly different from the words contained in a stoplist based on the most frequent words in a general usage corpus.

B. Searches completed using stoplists generated from an Air Force corpus will be significantly faster than searches done using a standard stoplist, and will produce the same, if not more accurate results.

C. A stoplist generated from an Air Force corpus will remove a significantly greater number of words, from a set of randomly selected queries, than a standard stoplist will.

Summary

This thesis is not intended to present an argument for the development of new or better search techniques. The main purpose is to develop and demonstrate a method of creating a stoplist that will improve the speed of current search techniques. The impetus for this thesis originated in a thesis by Capt David Snoddy. In his thesis Capt Snoddy recommended further research be done to create a stoplist tailored to the peculiarities of the usage of words with in the United States Air Force. Capt Snoddy used a very general corpus in thesis to classify records, and postulated a stoplist tailored to the unique language of the Air Force would provide better results in the application he developed (Snoddy, 1996:57).

Little to no research has been published on the specific topic of stoplist building or creation, and what has been published is very generic in nature. Chapter II will summarize the information currently published regarding stoplists and their creation. Chapter III will outline a simple method for creating a subject area stoplist based on techniques proposed for use in Chapter II, and outline how we intend to test the effectiveness of this stoplist. Chapter IV will summarize the development of a set of stoplists using the methodology constructed in Chapter III. Chapter V will discuss the gathering of information using our newly developed stoplists and the testing of our stoplists. Chapter V will also determine the effectiveness of the new stoplists from a quantitative and a qualitative standpoint. Chapter VI will discuss conclusions that may be drawn from the results of Chapter V, outline areas for further exploration within this topic, and provide some conclusions concerning the effectiveness and utility of stoplists.

II. Literature Review

Chapter Overview

The first research objective of this thesis is to summarize the pertinent literature pertaining to the research problem. As mentioned in Chapter I, search techniques and stoplists are a part of the Information Retrieval area of interest. As we will see, stoplists do not fill a very big part of this world. We will summarize the published material dealing directly with the creation of stoplists. The rest of the Chapter will be used to discuss index and abstract creation techniques; two related areas in the field of information retrieval that deal with the location of highly significant words rather than insignificant words. From the information we see in the creation of indexes and abstracts, we will build a simple method for creating a corpus specific stoplist.

Stoplists

It would appear stoplists are the unsung heroes of the world of information retrieval, but they are an important part of each information retrieval system. One of the simplest questions that should come to mind when reading this thesis, is why use a stoplist, what is the benefit in researching the construction and creation of such things? In their book, Information retrieval: data structures and algorithms, William B. Frakes and Ricardo Baeza-Yates (Frakes, 1992; 113) highlight several important points concerning the importance of stoplists. First,

> Stopwords may account for as much as 20 to 30 percent of the actual words within a single English document (Francis and Kucera 1982). Eliminating such words from consideration early in the automatic indexing speeds processing, saves huge amounts of space in indexes, and does not damage retrieval effectiveness.

This work, completed in 1990, was done without the knowledge of what the Internet would bring to the world in terms of information creation and access. Frakes and Baeza-Yates do not elaborate on the proper process, or suggested process for creation of a stoplist. They do make an important point concerning one of the most basic ways of creating stoplists. It is important to realize among a list of the 200 most commonly words occurring in English are words of significance such as "time," "war," "life," and "world." Indicating strict attention must be paid to words contained in a stoplist that may need to be considered for deletion due to their actual significance. This is a significant point to keep in mind when we are formulating our process of creating stoplists.

Actual documentation on a process for the creation of a stoplist is documented in a journal article by Christopher Fox. In Volume 24 of the Special Interest Group for Information Retrieval Mr. Fox, outlines his efforts in building a stoplist for use in a text analysis program (Fox, Fall 89:19). It is considered common practice to use the most frequently occurring words as measured from a large body of works written in the same language. The method used by Fox can be summarized in four simple steps:

1. Determine the frequency of words in a collection of written documents.
2. Limit list to size appropriate for application (establish cut-off point).
3. Remove frequently occurring but significant words.
4. Add words which missed cut-off point but are insignificant.

Fox adds an additional step because of the application for which he is developing the stoplist. This method, or technique, summarizes what is probably done in most applications, which is not to say it is incorrect or ineffective. The frequency of the words is gathered from a large collection of written English, referred to as a corpus, from a wide variety of subject areas. In the case of Mr. Fox's stoplist, the Brown Corpus is used as the source. The Brown Corpus is collection of 500 documents all published for the first time in 1961. The corpus contains approximately 1,014,000 words of running text from 16 areas of writing (Francis, 1982:5).

The article by Fox is the only one found on the stoplist creation. It does seem quite straightforward and rather uncomplicated. The most difficult part of the process outlined is in determining what words to not include in the stoplist and what should be added to the stoplist. Selecting words that are important and occur very frequently could consume quite a bit of time.

Two important fields of work in the area of information retrieval are the automated creation of indexes and abstracts. These two subject areas concentrate on extracting the most important pieces of information from a document. Many methods are used to determine which words are most significant and will best convey the essence of a document. Using the frequency of a word within a document and its position within a sentence is one method seen quite often. If it is possible to determine what is significant in a document it should also be possible to determine what is insignificant within the same document using very similar techniques.

Indexing and Abstracting

The process of indexing and creating abstracts of literary works has been done for quite some time. It has always been a very laborious, slow and intellectually wasteful task performed by librarians. Researchers and scientists, since the introduction of modern computers, have tried to apply computer technology to any task performed by humans. Researchers have also been using computers to explore the importance of the frequency of words within a document and their relative position in a sentence tried to apply this to the automatic creation of indexes and abstracts.

Many early efforts in automatic abstract and index creation centered on combinations of the frequency of a word and the location of a word within a sentence. A graph of the product of the frequency, *f*, of a word in a given position in a text and the rank order, *r*, that is the order of their frequency of

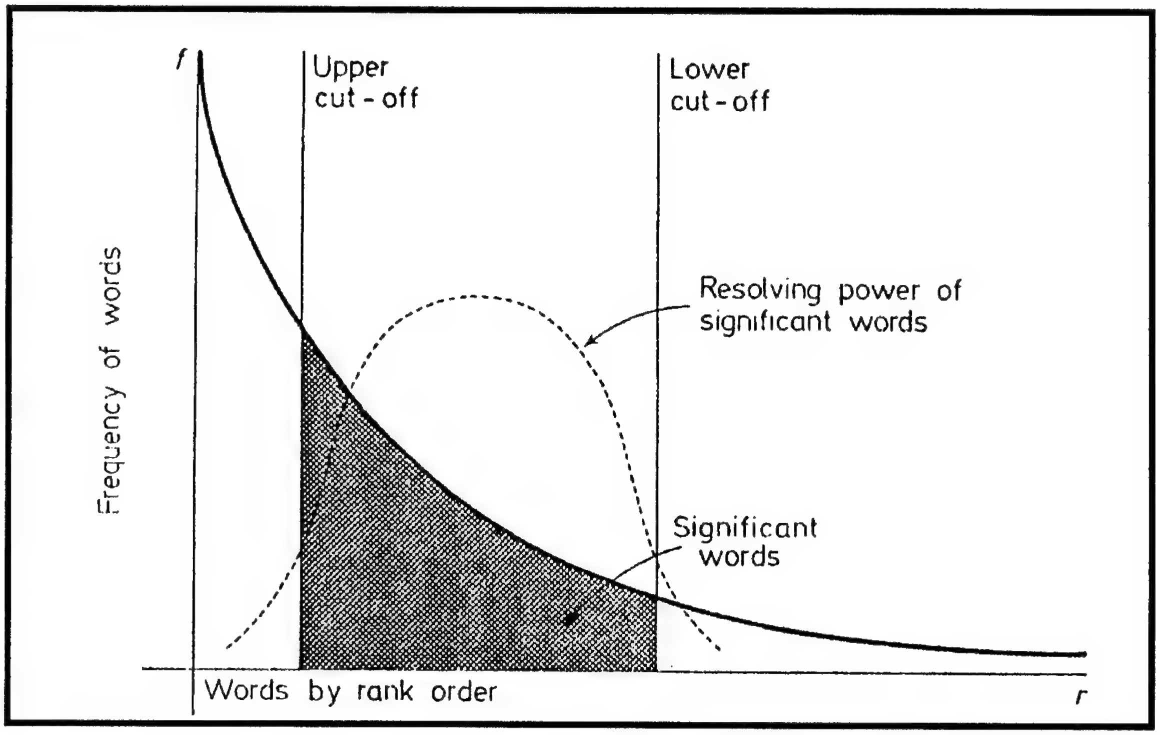


Figure 4. Graph of f (frequency) versus the rank order

occurrence, will yield a graph similar to Figure 4. Words falling between the upper and lower cut-offs would be very effective as part of an index or abstract. Notice the significant words are not determined entirely by their frequency, but as a function of frequency and their rank order. Words that occurred too frequently were considered common, and those below the lower cut-off rare and insignificant (van Rijsbergen, 1979:14). It is conceivable to adopt this idea to identify the insignificant words in a corpus. A drawback of this approach would be the amount of information to be collected to calculate this information. A tabulation of the occurrences of each word in a corpus could be rather large; while a tabulation of the number of times a word occurs in a specific position in a sentence would be quite a massive undertaking.

In 1959 H.P. Edmundson and R.E. Wyllys, of the Planning and Research Corporation, compiled a report summarizing several methods of creating automatic abstracts and indexes. It was "understood that a frequency count of the significant words of a document can serve to isolate the special vocabulary used to convey information in any particular realm of discourse." (Edmundson, 1959:34) Our efforts are concerned with words of little significance, and we need to find a way to extend this theory to a collection of writing, or corpus. One of the methods summarized focused on the idea that "a word's information should vary inversely with its frequency rather than directly, its lower probability evidencing greater selectivity or deliberation in its use." When a word occurs more frequently in a document (or a collection of documents) than in general usage (or a collection of general writings) it would carry more information. Comparing the relative frequency of a word within a document , *f*, with the relative frequency of the same word in general usage, *r*, can provide an indication of the word's value in the preparation of an abstract entry (Edmundson, 1959:36-37).

This theory could be implemented as follows:

Begin by determining the total number of words, w , in document d, symbolized by $N_d = \sum_{\text{all of d}} N_{wd}$, the total number of times each word w appears in document d, symbolized by , $N_{wd} = \sum_{\text{all of d}} \text{w}$, and the frequency of each word in the document d, symbolized by $f_{wd} = \frac{N_{wd}}{N_d}$.

Similar computations must be accomplished for each word with respect to its use in general usage. We substitute r for f , and c for d . The transformation from d to c indicates a shift from a single document to a collection of documents. It is now possible to obtain a comparison, a measure of significance, between the usage of a word in a specific document and the usage of the word in general.

The authors identify four specific measures expressing the significance of a word s_{wd} by comparing the frequency of the word in a document to its frequency in general use. The four measures are shown in figure 5.

$$s_1 = f - r \qquad s_3 = \frac{f}{f + r}$$

$$s_2 = \frac{f}{r} \qquad s_4 = \log \frac{f}{r}$$

Figure 5. Measures of significance

Each of these measures indicates a different value for each word. Graphing the values of s versus f would give a graphical representation of the significance of each word with respect to its frequency in general use. In general, if you were to graph each of the functions for all the words within a document, you should obtain graphs similar to those in figure 6.

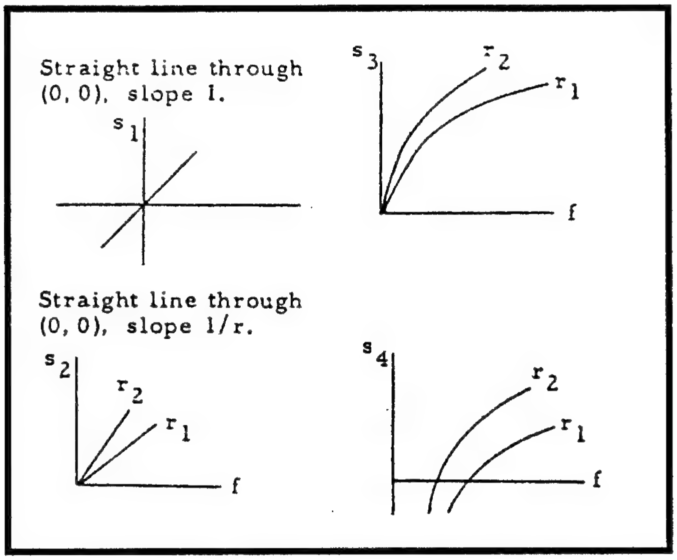


Figure 6. Graphs of measures of significance

The graphs of the four measures of significance should help in determining the relative significance of each word within a document. The author notes;

> Whichever of these simultaneously monotonic functions is chosen, it is clear that defining significance in terms of the contrast between frequency in a document and in general usage would give low significance both to the normally rare words which occur rarely in the document and to common words used frequently within the document, while giving high significance to normally rare words used frequently.

So the determination of the insignificant words within a document could be done using this method if it was possible to obtain the general usage information for the words within a document.

It should be possible to compare the use of words in a subject area corpus to their use in a general corpus and identify the insignificant words in the subject area corpus.

Sources Searched

The search for information pertaining to the creation of stoplists was quite extensive and involved the use of multiple information gathering tools. The first source of information used was FIRST SEARCH's Engineering and Technology area's ArticleFirst and INSPEC databases. The two databases were searched using the terms "stoplist," "stopword," and "negative dictionary." Although each of the databases indicated articles were found with these words in their abstract closer review indicated they merely mentioned the fact a stoplist was used in a specific application. None of the articles dealt with the creation of a stoplist.

World wide web meta search engines were used to look for material stored on publicly accessible web sights. A meta search engine submits a search query to a large list of World Wide Web (WWW) search engines simultaneously, and in the format used by each of the specific search engines. This allows searches to be carried out in parallel instead of one at a time. Dogpile and Metacrawler, two of the more powerful meta search engines, were used to submit queries such as "stoplist," "stopword," and "negative dictionary." Documents returned in response to these queries fell into two categories. Responses were either documents mentioning the use of a stoplist or negative dictionary without discussing how the stoplist was created, or they were documents dealing with stoplists for organs and organ music.

The Defense Technical Information Center's database of published research articles, obtainable on CD-ROM in the AFIT library, was also searched. Subject, keyword, and abstract keyword searches were done using the words "stoplist," "stopword," and "negative dictionary." Again, the occurrence of these words was in reference to the use of a stoplist not its creation.

Summary

Although we have been manipulating documents for quite sometime, and even using stoplists in the process, we seem to have left the creation of stoplists to the creativity of programmers and text processors. Of the literature examined we do not see one clear anything identifying a process for creating a subject area stoplist. We also do not see any literature supporting a claim for improved efficiency or accuracy when subject specific stoplist is used.

We do find there are gains to be made from the use of a stoplist in reducing the size of indexes and therefore the storage requirements for systems using indexes. It is reasonable to expect a performance gain for any world wide web or database system using an index when a stoplist is used to limit the size of its index.

Stoplist creation is an area of little research. The single article we examined pointed to a very direct method based on the use of the words occurring most frequently in general usage. Other articles suggest it may be possible to determine more accurately identify less significant words in a specific subject area by comparing their usage rate in the subject area to their usage in general texts. The existence of such words is not as intuitive as it may seem. We do not see any literature suggesting these two ideas be combined as a way of identifying words of low significance.

My proposal is to apply the methods discussed by Edmundson and Wyllys for the determination of significance measures for words and applying them to a corpus. Using the information on the frequency of individual words, gathered from an electronic collection of Air Force documents, we will determine which measure provides the best indication of a word's insignificance. It is logical to compare the frequency of a word in a subject area to its frequency in general usage to determine insignificant words in a given area. This differs from using just the rank ordered frequency of words from a given area as a source for a subject area stoplist. Calculating the four measures of significance and plotting their values for an entire corpus should clearly show what words carry significance and what words do not.

III. Methodology

Chapter Overview

The second objective of this thesis is to develop and test a simple (2-3 step) process for creating a stoplist tailored to a subject area. To satisfy this objective, we will outline a process based on methods discussed in the literature review. Then we will detail the steps used to create a stoplist based on a collection of Air Force documents (hereafter referred to as a corpus) using the simple process we outline. Next we will outline the testing process used to obtain information to determine the effectiveness and efficiency of an Air Force specific stoplist using the Air Force's homepage. We will end the chapter by describing the quantitative and qualitative analysis methods to be used in the Analysis and Results section of the thesis.

Process for Creating a Subject Area Stoplist

Our literature review shows there is not an overwhelming amount of published information on methods for creating stoplists. Because of the limited amount of published information we do not see any literature on a method for creating a stoplist for a subject area. The method described by Fox (Fox, 1989:19-20) outlines the basic steps involved, and there is no reason to disregard these steps since they do seem logical. The one area needing examination is how to determine if a word is insignificant and should be part of a stoplist. Fox bases selection solely upon the frequency of a word in general usage. This premise may be appropriate for stoplists used in applications with information from general backgrounds. An example of such an application might be a search engine for a CD-ROM based encyclopedia or an on-line search engine for a public library. These examples would be searching documents or databases with text from many areas of study. The opposite end of the spectrum, and the focus of this research, is an application focusing on a specific subject area. Appropriate examples would be a CD-ROM version of Jane's all the Worlds' Aircraft, a medical research database, or a Central Intelligence Agency database focused on foreign made spacecraft. These examples, although they contain millions of words, vary greatly in the specific words that occur in them compared to the specific words that occur in the previous, (more generic) examples. It therefore stands to reason that a stoplist intended to be used in an application that deals with a specific subject area should use a stoplist created for the specific application.

It is possible that a measure other than a word's actual frequency may better determine it's worth as a stopword. van Rijsenberg's theory of the resolving power of significant words implies there is a definite relationship between the frequency of a word's usage and it's resolving power. This theory is aimed mainly at the preparation of indexes and abstracts and has limited applicability to this effort, but is still important. It is important because it helps point out that frequency can be useful in determining a word's power or significance.

To illustrate, consider an example in which we examine words that occur frequently within a given subject area (e.g., the area of personal computers). We might find the word "computer" is the most commonly occurring word while the word "house" occurs two times. Because of this, we would probably find the word "computer" to be of little or use in locating pertinent information in a database of information on personal computers. Examining a collection of written English, such as the Brown Corpus, would show the word "computer" occurs 18 times out of a total of 1,014,000 word occurrences, and is ranked as the 4,711th most frequently occurring word of over 36,000 unique words. As a comparison, the word "house" occurs 662 times, and is ranked as 125th most frequently occurring. Thus, it seems quite clear that the frequency of a word's usage is going to differ depending on whether you are using a corpus of general usage or a subject corpus. Examining the frequency of a word in both corpuses would seem to provide us with a more effective means of determining a word's significance (i.e., suitability as a stopword) for a specific application.

Edmundson and Wyllys' work uses a word's frequency, in both general usage and in a subject area, to determine it's significance. The four measures of significance, shown in figure 5, provide a means of assessing the significance, or insignificance, of a word. It is for this reason we will use their measures of significance and the graphing techniques they describe to determine which words will be used in our Air Force stoplist.

Our process is outlined as follows

1. Determine the frequency of words in general corpus.
2. Determine the frequency of words in subject corpus.
3. Determine measure of significance.
4. Graph each of the measures of significance.
5. Limit list to size appropriate for application (establish cut-off point).
6. Remove frequently occurring but significant words.
7. Add words which missed cut-off point but are insignificant.

This process closely follows the process outlined by Fox while using the measures of significance as the determining factor of a word's significance rather than its frequency alone.

We have created a process that should allow us to easily identify words of low significance in terms of their suitability for searching. One item not mentioned before is the readability of the graphs from step four. The process this technique was taken from was designed around the creation of abstracts of articles. An article is generally much smaller than a corpus and contains a significantly smaller number of unique words. This fact may lead to a graph that is very difficult to interpret. We will have to determine if this is in fact true in Chapter IV of the thesis.

Stoplist Creation

Since we have established a procedure to create a stoplist we will begin a description of an implementation of these steps for the United States Air Force. The steps above will be followed and we will also outline intermediate steps considered relevant to the process.

Select Corpuses. The first step is the selection of two corpuses, one to represent the general usage of words and one to represent word usage in the Air Force. The literature reviewed does not outline a specific technique for selecting a corpus or how to produce one if it is necessary. Professionals have compiled a quite large collection of corpuses in various languages and in many different formats. Corpuses of English language exist and are use quite heavily in the field of linguistics. The following paragraphs will outline some consideration when selecting which corpuses to use and what documents to select when creating a corpus.

General Usage Corpus. As the representation of a larger collection of English usage the selection of the general usage corpus is very important. Unfortunately Edmundsun and Wyllys do not specify the design of the general usage nor do they make any recommendation as to its composition. We can establish basic guidelines to be used when evaluating prospective corpuses.

- The size of the general usage corpus should be relatively larger than the size of the Air Force corpus. Although we are dealing with the frequency of the words involved, a larger corpus will provide a broader range of words for analysis.
- The corpus should not concentrate its selection of materials from a specific area of interest or type of document (i.e., it should contain text from philosophical authors as well as commentaries by average citizens).
- Information should be presented or provided in a format that is simple to work with and make determining the word frequencies as easy as possible. A corpus stored in a proprietary format, requiring a specific reader or decoder, would produce necessary complications.
- A concern in a research effort is the cost of using or obtaining copy of the corpus. Creating a corpus can be very time-consuming and corpus creators often charge for the privilege of using their work. This may not be a great concern for a funded effort.

The issues listed above are not supported through extensive research, but are based on common sense and a limited amount of information on the construction, composition and general characteristics of corpuses in general. Given the variety of corpuses available it should not be too difficult to find one that fits most of the needs listed above.

Air Force Corpus. Most of the concerns outlined for the selection of the general usage corpus will apply just as equally when selecting the Air Force corpus we will use. The size of the corpus should not be as much of a concern as it is in the selection of the general usage corpus. It does however need to be representative of the many areas of interest and specialty within the Air Force. So it is important the corpus selected have a breadth and depth of diversity. A corpus in a familiar format will be easier to work with and processing will be quicker. The cost of the corpus will probably not be a concern, but the availability of such a corpus may be. We will have to look at all the available options and evaluate each one of them individually.

Determine Word Frequencies

The frequency of the words within each of the corpuses is key to determining the measures of significance for each word. Once each of the corpuses is identified it will be necessary to develop a program to count the number of times each of the words appears in each of the corpuses. Several text analysis packages will perform this function and are available as either shareware or freeware. These programs are unfortunately designed to work with small amounts of information and not with corpuses. Frequency counts for corpuses are usually custom made programs operating on mainframe computers and are not freely distributed. To perform this task the same approach and a program was written to count the words in both of the corpuses. The following paragraph outlines the program that was written.

FREQOUT Program. Preparation of the program to count the number of times each word in the Air Force corpus occurred could now be pursued more vigorously. The program itself would be written as a Perl script. A form of programming used in the UNIX environment to execute system level commands in an automated function. The script would execute the commands to open all the files constituting the corpus, identify the words within it, and count the number of times each one of them appeared. It would maintain a file that calculated the total number of times each of the words appeared. The appearance of contractions was of concern because of the difficulty in writing the script so it would recognize when a word started and ended. The simple definition is to find a collection of continuous letters with a space at the beginning and the end of them. This would require treating all punctuation marks, including the apostrophe, as spaces. Converting the apostrophes to spaces would result in contractions being truncated, and a word such as "couldn't" would be counted as the words "couldn" and "t." Considering the amount of time required to write the script to properly count the contractions, it was decided to ignore the contractions and treat punctuation marks as spaces. The final script, shown in Appendix A, reads the files from the directory the script is located in, and writes the output to a separate file in the same directory. The output file lists the words, the total number of times each word appeared, and the total number of words counted in all the files. An example of the output is shown in table 1 below.

Table 1. Excerpt from output of FREQUOUT script

WORD	COUNT	WORD	COUNT	WORD	COUNT
a	130141	aadc	2	zurich	1
aa	120	aads	2	zv	1
aacs	10	aae	1	zva	1
aad	59	.....	.....	zvu	1

Find Word Overlap

Restricting further analysis to words appearing in both the Air Force Corpus and the BNC will allow us to analyze only those words that are correctly spelled and entries that are words rather than acronyms. Only commonly accepted acronyms, such as AF or AFB would be part of any further analysis. This would have an unrelated advantage of reducing the size of the files used to store the words to be analyzed. An important consideration when using a list containing hundreds of thousands of words. Additionally, measures of significance S2 and S4 would not be calculated in cases where a word appeared in the Air Force corpus but not in the BNC. The measures of significance for words in this situation would be indeterminate since the frequency of the word in the BNC, r , would be zero. The calculations for both S2 and S4 involve dividing by r. We are now ready to compute the frequencies of the words in each of the corpuses.

Calculating Word Frequencies. The number of words in the overlap area will indicate what software program will be used from this point forward. If there are less than 4,000 words in common between the two corpuses, we will use Microsoft's Excel Spreadsheet program. This stems from the fact that Excel will not graph more than 4,000 points on any one graph. If the total number of words is more than 4,000 we will need to use a program designed more for mathematics such as MATHCAD. MATHCAD will allow users to import files containing more elements than Microsoft's EXCEL can accommodate.

With files containing the counts of each word in both of the corpuses we will determine the frequency with which each word appears. Word frequencies are determined by dividing the number of times an individual word appears by the total number of words within the entire corpus. It will be important to maintain a significant number of decimal places of accuracy and all calculations will be carried out to 12 decimal places. With programs such as Microsoft Excel and MATHCAD this will be simple to do.

Calculate Measures of Significance

Our next step will be to calculate the four measures of significance for each of the words we are now analyzing. The calculations for each of the measures of significance are basic and will be easy to compute. Either program, MATHCAD or Excel, should provide the answers in manner that will facilitate easy graphing of the resulting measures of significance.

Graph Measures of Significance

Our next step will be to graph the computed measures of significance to see if we can determine a clear and definite cut-off point through visual inspection of the plots. According to Edmundson and Wyllys, our graphs should look very similar to those shown in Figure 5 (Edmundson, 1958:34). We should be able to determine exactly where the insignificant words end, and tell what word to cut the stoplist off at using the indicated measure of significance to determine the actual cut-off. We will select a cut-off for each of the four measures of significance and determine the words to be included in each of the four stoplists.

Limit Stoplist Size

Determining the number of words to use in a stoplist is not an exact science. Throughout the literature reviewed the overwhelming factor in stoplist size determination was the effect upon the search engine being used. Other than this concern, Fox constructs his stoplist by establishing an occurrence rate, looking only at words occurring more than 300 times in a corpus, and then refining his list (Fox, 1989:21). This stoplist includes 425 words and was independent of the application being used. Snoddy used this stoplist as part of his RACS program with the addition of two words he felt needed to be included (Snoddy, 1996:28). For our application, we will need to consider the number of words left in consideration after finding which words are in common between the final Air Force corpus and the BNC.

Review Stoplist

Although we have compiled our four stoplists we must ensure our procedure has not left out any words that should have been intuitively included in the list or included a word that may be considered too significant to be included in a stoplist.

Remove Significant Words. Some words may be deemed insignificant through a calculation, but are essential to locating information within an application. In the case of the Air Force Corpus it would be very difficult if we were to allow the word "Blackbird" or the name of an Air Force installation, such as "Offutt," to be included in a stoplist. This is not a process which can be easily documented, but is based more on familiarity with the subject area and the importance of each word.

Add Insignificant Words. It is also important to scan the stoplists and ensure words you would expect to see are present. This process may be difficult because of the size of the stoplists, the number of stoplists involved, and the number of words in the Air Force corpus. This is a check of the words in the stoplist to make sure most of the words you would assume to be present are in fact there. Words common within the Air Force corpus should be looked for, and if they are not found, the calculation for their measure of significance must be examined for accuracy. A final sanity check is always important before implementing it in a functioning system.

Testing

At this stage we should have four stoplists, one for each of the four measures of significance, and feel confident they contain words considered to be insignificant through one measure or another. Creating a more efficient stoplist, one that operates quicker, is an underlying goal of this thesis, but it is inconsequential if the new stoplist does not provide accurate results.

To determine if these stoplists are effective and efficient, it is necessary to test them in an actual application. For this purpose, we will use the query function on the Air Force's homepage and a set of randomly selected queries. Each of the four stoplists we have created will be tested with all the test queries, and compared with information obtained when the test queries are submitted while a baseline stoplist is used. The baseline stoplist will be taken from the most frequently occurring words from the general usage corpus.

Along with the queries we have selected, we must collect information that will help support or reject each of the hypotheses made in Chapter I. Although the Air Force homepage was selected as test application early on, it is important to show this is a reasonable choice for our purposes.

Suitability of Air Force Homepage as Test Application

Our stoplists were created from a corpus consisting of Air Force documents so we must limit our testing to a system or site dealing with Air Force material, articles or text exclusively. Many sites fitting this description are available and are easily accessible. A major concern is to find a site willing to, and capable of, allowing an outside party to provide several stoplists for use on their site. This concern eliminated most databases in the Air Force, since operating them with a stoplist, which is experimental at best, would be rather unwise and risky. The Air Force's homepage was considered an excellent candidate because of its size and the design of the search engine it used. The Air Force homepage, while a specific topic area, also contained information from many sources and on many different subjects. Because this homepage operated as shown in Figure 1 it was an excellent candidate to help prove or disprove the first two research hypotheses.

The Air Force homepage is operated by the Defense Technical Information Center (DTIC) at Fort Belvoir VA. DTIC's Information Support Directorate agreed to provide assistance as needed and to allow the homepage to be indexed using the four stoplists. Information needed to compute statistical measures for use in Chapter IV was also provided, including system load statistics, search engine logs, and index file sizes. All of these items would be needed when performing the quantitative and qualitative analysis. The logs from the search engine were very essential in determining what queries would be used during the actual testing. It is because of this sites broad range of information, and the administrator's willingness to provide the needed assistance that it was selected as the test application. Having selected a test application it would now be easier to obtain sample from which to select our test queries.

Test Queries

To maintain objectivity, and ensure the results of the research effort were acceptable, it was determined that queries entered by other users would be used for the majority of the test queries. The individual query terms submitted by unknown users will be extracted from search engine logs for three consecutive days. Next, the queries will be placed in a spread sheet, and each of them assigned a random number. The queries will then be reordered according to the newly assigned random number. Next, the total number of queries will be divided by the number of test queries desired to determine how often to select a query from the randomly listed ones, this number is called our offset. Beginning at a randomly selected number between one and our offset we will begin selecting queries that occur at an interval equal to the offset, determined above. As an example, if we collect 6,432 queries, our offset will be 129. If our random number is 65, then queries 65, 194, 323 etc. will be chosen as our test queries. It was decided 50 queries would be sufficient to provide a good sampling of results from each of the stoplists. Using a sample size of 50 allows us to assume the underlying population of the sample we have selected will be approximately normal. Assuming our sample is normal will allow us to meet most of the assumptions required to use parametric tests.

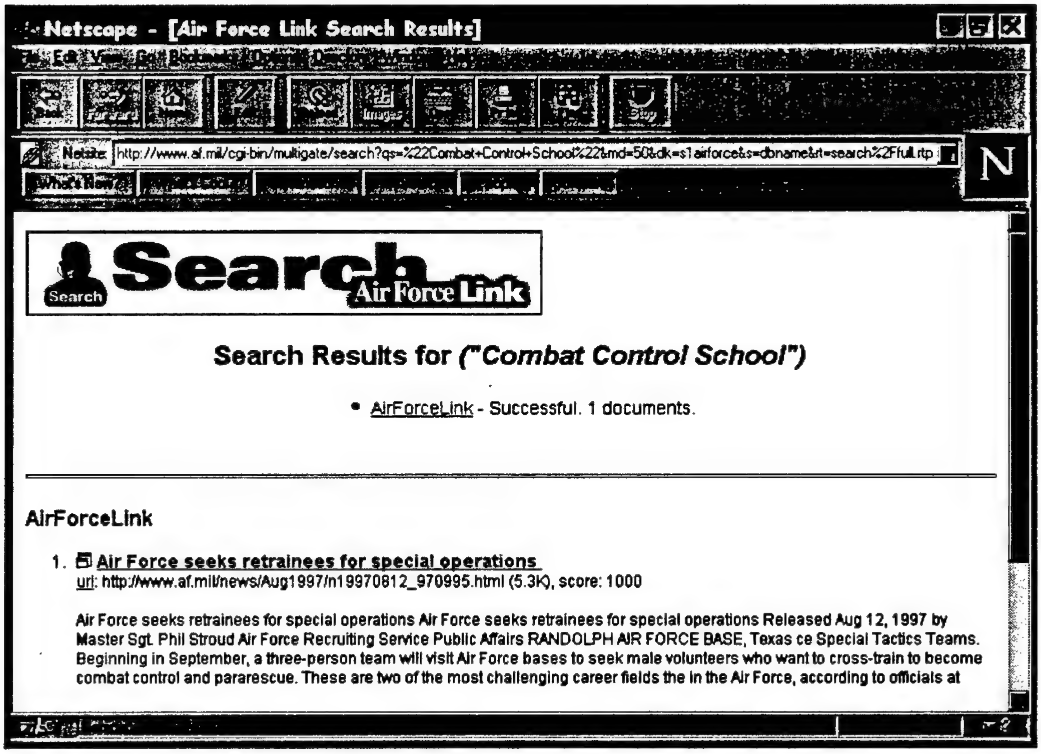
Besides the 50 random user queries, it was decided to add an additional 30 queries taken directly from items located on the Air Force homepage. Items were taken from all areas available on the homepage and from various parts of different articles. The excerpts were stored as exact quotations from the documents and would be submitted to the search engine as direct quotations. The intent was to have a mix of queries to provide a diverse set of inputs.

There were also other considerations made when trying to determine how many queries to use. The amount of time required to enter each of the queries, save the results from the query, record the number of items returned by the search engine, and to determine the amount of time used by the search engine during each query were also major concerns. Obtaining a very robust set of queries is of little use when there is not enough time to process them all or to evaluate the results from them. A sample size of 80 is sufficiently large to allow us to make some assumptions about how well our sample represents the underlying population it represents. Our population is made up of all the queries submitted to the Air Force homepage. The number of words contained in each of these queries is unique, but using a random sample of 80 queries allows us to use this as a representative sample of our unknown population.

Since the Air Force homepage's search engine uses the "OR" function for multiple word queries, some of the queries may need to be odified. As an example, if a user submits query of Samuel C. Robbins we will change this to "Samuel C. Robbins." Adding quotation marks will limit the search to those documents that contain the entire query instead of any document that mentions any of the three words in the original query. Those queries already containing search qualifiers, such as "AND," or "NEAR" were not modified.

Gather Information

Before testing a stoplist, the Air Force homepage will be reindexed using the stoplist being tested. Reindexing will remove all the stopwords from the homepage's index, an example of which is shown in Figure 1. Each of the 80 queries will be submitted to the Air Force homepage's search engine, and three pieces of information will be recorded for each of the queries as they are processed. Two pieces of information will be take from the search engine results page, the number of successful documents (or hits), and the summary information for each of the documents. An example of the results page is shown in figure 7 below. The search engine results page displays information in an easy to interpret format, and provides information such as the query as it was entered, number of "hits," the uniform resource locator (url) for each of the items the query was found in, and a ten line summary of each document.



http://www.af.mil/cgi-bin/multigate/search?qs=%22Combat+Control+School%22&md=50&dk=s1airforce&s=dbname&rt=search%2Ffull.rtp

Search Results for *("Combat Control School")*

- AirForceLink - Successful. 1 documents.

AirForceLink

1. **Air Force seeks retrainees for special operations**
url: http://www.af.mil/news/Aug1997/n19970812_970995.html (5.3K), score: 1000

Air Force seeks retrainees for special operations Air Force seeks retrainees for special operations Released Aug 12, 1997 by Master Sgt. Phil Stroud Air Force Recruiting Service Public Affairs RANDOLPH AIR FORCE BASE, Texas ce Special Tactics Teams. Beginning in September, a three-person team will visit Air Force bases to seek male volunteers who want to cross-train to become combat control and pararescue. These are two of the most challenging career fields the in the Air Force, according to officials at

Figure 7. Search Engine Results screen from the Air Force homepage

Number of Successful Documents Returned. The number of successful documents returned by the search engine results page will be recorded in a spreadsheet for each of the queries submitted. The number of successful documents will suggest how much each of the stoplists has limited the scope of the search with respect to the other stoplists. An "*" in our spreadsheet will indicate there was an error of some type encountered by the search server. A "-" will indicate 200 or more documents returned for that search. The Air Force homepage search engine will not return more than 200 documents for any query, which poses a problem for our purposes. It is not possible to determine if the actual number of documents was 201 or 2,001. A large difference such as this may cause large variances in the amount of time or memory used to implement the search. Not being able to determine the exact number of documents returned makes statistical analysis of these queries unreasonable leading to the placement of the "-" in all the data tables for these situations.

Information Returned. The contents of the search engines results page will be stored to allow the results from each of the queries to be compared with results when another stoplist was used. Recording the contents of the results page will allow the retrieval of the specific documents returned for each of the queries submitted.

Search Engine Log. The search engine's log records the time each of the steps in the search process is executed. The log can only provide information accurate to the second. The information in the log will be used to determine how much time each of the queries took to execute. An example of the log is shown in Figure eight.

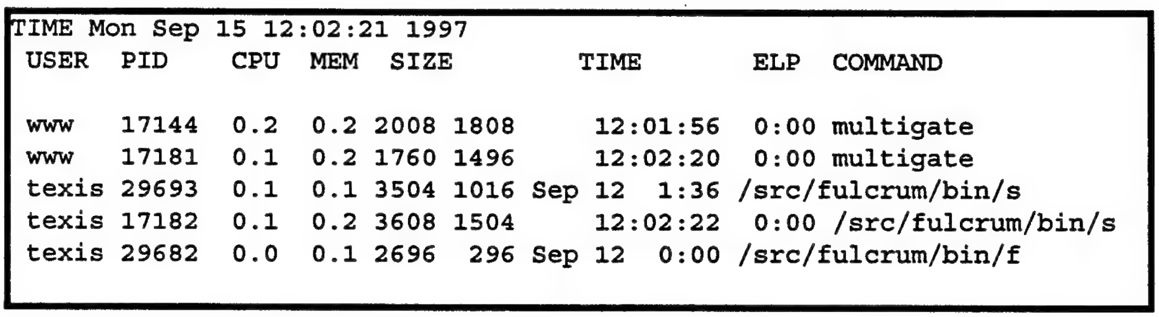
```
17182: 0: Sep 15 12:02:25 1997: 100: init message: user 131.84.1.31 password <none> client
MultiGate
17182: 1: Sep 15 12:02:28 1997: 3: search database: "airforce" seed words: "(scouts)"
17182: 2: Sep 15 12:02:28 1997: 100: SQLExecDirect: SET RELEVANCE_METHOD 'F2:4';
17182: 3: Sep 15 12:02:28 1997: 100: checking access for client 131.84.1.31 to database airforce
17182: 4: Sep 15 12:02:28 1997: 100: SQLExecDirect: SELECT relevance() as score, TITLE,
FIRSTTEN, new_url, FT_CID, FT_FLIST, FT_FORMAT, FT_ORIGINAL_SIZE          airforce
WHERE FT_TEXT CONTAINS 'scouts' ORDER BY score desc
17182: 5: Sep 15 12:02:28 1997: 100: result rows = 27, columns = 8
17182: 6: Sep 15 12:02:28 1997: 4: found 27 hits
17182: 7: Sep 15 12:02:30 1997: 2: closing connection;
17182: 8: Sep 15 12:02:30 1997: 2: done handling client
29693: 23141: Sep 15 12:02:30 1997: 100: child PID = 17197
```

Figure 8. Search Engine Log Example

The log shows the query found 27 matching documents, and it created a new process, process # 17197, to generate the results to the screen for the user. The amount of time required to process each query will be determined by subtracting the time the search was completed, signified by the statement "done handling client," from the time the query was received by the search engine, indicated by the statement "init message."

System load Statistics. The server the search engine is running on creates another log that records system load statistics at regular intervals. The system load statistics indicate how much of the server's physical resources, such as memory and processor capacity, are in use at a specific time. This Information will be used to account for extreme variations in the amount of time it takes a query to be processed.

The server the Air Force homepage is operated on records system load information every five seconds, an example of the log is shown in Figure 9. The log identifies the type of process running, the PID, the percentage of the total system memory being used by the process, the percentage of the total central



```
TIME Mon Sep 15 12:02:21 1997
 USER  PID    CPU  MEM  SIZE         TIME       ELP  COMMAND

 www   17144  0.2  0.2 2008 1808      12:01:56  0:00 multigate
 www   17181  0.1  0.2 1760 1496      12:02:20  0:00 multigate
 texis 29693  0.1  0.1 3504 1016 Sep 12  1:36 /src/fulcrum/bin/s
 texis 17182  0.1  0.2 3608 1504      12:02:22  0:00 /src/fulcrum/bin/s
 texis 29682  0.0  0.1 2696  296 Sep 12  0:00 /src/fulcrum/bin/f
```

Figure 9. Example of server usage statistics

processing unit (CPU) being used by the process and the time the process started. Figure nine shows system load statistics recorded on September15 1997 at 12:02:21. The first process listed is a WWW process consuming 0.2 percent of the system's total memory and 0.2 percent of the system's total processing capacity. With the PID for a process, it is possible to determine how much of the system's resources a specific job is using. Where a process is listed in more than one interval, the information from the latest interval will be recorded.

Analysis Techniques

The efficiency and effectiveness of the stoplists we have created are of importance to this research paper. We will now examine how to determine the effectiveness and efficiency of the stoplist how we will determine if they support the hypotheses in Chapter I.

Three separate quantitative hypotheses were made regarding the stoplists and their effect on search engine efficiency. Hypothesis B also made a qualitative claim. We will examine each of these hypotheses independently and describe a method for testing each one. We will test each of the hypotheses with each one of the four stoplists we create, and select one of the stoplists that best supports all the hypotheses. Each of the areas of analysis is described along with the appropriate conditions for failing to accept the hypothesis. Any one of the stoplists may prove viable in not rejecting the hypothesis, but can be considered as significant unless it is viable under all the hypotheses.

Quantitative Analysis

Each of the hypotheses involves quantitative analysis of one type or another. Each of the hypotheses will be examined individually, and a method of testing each one will be outlined.

Hypothesis A Differences in stoplist Contents

Hypothesis A was meant to determine if there was a significant difference in the words contained in the Air Force stoplists compared to the words contained in the stoplist generated from the general usage corpus. This should show if there is a difference in words used in Air Force writing and communications.

We will determine the number of words in common between each of the stoplists, and divide by the number of words in the general usage corpus. This number will represent the percentage of overlap between two stoplists. Significance will be determined based on all the overlap percentages. Attaching statistical significance to this measure is not reasonable and will be used as an indicator, not a determining factor, in selecting which of the stoplists performs the best.

Hypothesis B Search Times

This hypothesis goes directly to showing an Air Force specific stoplist will decrease the amount of time it takes to perform a search. Using the data we gathered on the time to complete each of the queries we will use a paired differences t-test (McClave, 1994;420-424), using an α of 0.05, to determine if any of the stoplists have an average query completion time faster than the BNC stoplist. Using a paired differences t-test allows us to compensate for the lack of independence between S1, S2, S3, S4 and BNC. S1, S2, S3 and S4 are dependent on both AF and BNC and make direct comparisons, such as a difference of means test, unusable. We should be able to determine if any of our Air Force corpus generated stoplists allows the search engine to complete all the queries faster.

We will limit our data set to those queries where there is data present for all the stoplists. In other words, if there were more than 200 results returned for a specific query under just one of the stoplists, such as query number 69, all the data relating to the times for that query will be removed from the analysis. This will be done to facilitate the use of the paired difference t-test, and to maintain consistency in the calculations. We will limit our data in this manner for any analysis which requires the use of a paired differences t-test.

Hypothesis C Number of Words Stoplisted from Random Queries

One our basic assumption is the fewer words being searched for by a search engine the faster the results will be returned. A comparison of the number of words, from a random set of queries, stoplisted by each of the stoplists will provide an indication if one of them stops significantly more words than the others.

To determine this we will calculate the total number of words in the set of random queries. Then the queries will have the stopwords removed from them using each of the four Air Force stoplists and the general usage stoplist. The number of words stoplisted will be determined by subtracting the number of words remaining from the original total number of words. We will then use a paired differences T-test, using an α of 0.05, to determine if any of the stoplists stops significantly more words than any of the other stoplists. Using a paired difference experiment removes the variability due to the individaul queries (McClave, 1994;420).

Qualitative Analysis

Hypothesis B asserts an Air Force stoplist will be faster than the general corpus stoplist while providing the same if not more accurate results. It is essential we examine this hypothesis since it does little good to obtain our results quicker if they are not the results we are looking for. We must determine if our stoplist is effective in addition to being efficient. However, determining if two sets of results are comparable is not a simple mathematical task. Because of the number of queries and stoplists involved it is necessary to develop a systematic method of evaluating if the information returned is comparable.

Avoiding Threats to Validity

The evaluation of qualitative information has some inherent threats to validity. Any research conducted must take care to acknowledge the sources of these threats and determine how to best counter them. In determining which queries will be examined we must be sure to avoid one of two broad validity threats, researcher bias. Researcher bias would result from imposing preconceptions, theories or values on the selection of the queries (Maxwell, 1996:90). To counter this threat we will adhere to an explicit selection criterion designed to provide results reflective of the information examined, and to eliminate selection of data which conforms to expected criteria or data which stands out.

Selection of Results to Analyze

To accomplish our qualitative analysis we will limit our examination to a comparison of the results from two stoplists. One set of results will be from our baseline stoplist, the BNC, and the second will be from what we will call the best stoplist. In the preceding paragraphs we outlined three methods for comparing the efficiency of the stoplists we will create. We will make a quantitative assessment of these results and select the one stoplist that appears to present the best results. Selecting a best performer will allow us to perform a more thorough examination of the results it returned on each of the queries and compare them to the results returned when the general usage stoplist was used.

In the quantitative analysis section we outlined three comparison methods designed to determine which stoplist is the most efficient in each of the specific areas addressed. Of these comparisons, those outlined to test hypotheses B and C can both point out which stoplist is the best. The test comparison for hypothesis A would not necessarily help point to a stoplist that is more effective than the others, just the one which is most different. The stoplist that is deemed to be the most efficient in each of the areas will be assigned ten points towards selection as the best stoplist. The remainder of the stoplists will be given 20 points for second place, 30 points for thirds place, and 40 points for fourth place. The stoplist with the lowest cumulative score will be our best stoplist. If there is no way of determining, from the results of the statistical test, which stoplist performed better, all the available point will be divided among the stoplists equally for that test. Having outlined the process for determining the best stoplist we can now determine how we will compare the results between the two stoplists.

Method for Analyzing Results

Because of the large number of comparisons possible we will only examine those queries in which the number of documents returned was no more than 30 for either stoplist. Looking at queries that returned 30 or less documents should provide an acceptable representation of the overall population of query results. Examining queries with 30 or less documents will also limit the time devoted to qualitative analysis and allow more time to be spent in other areas.

The results obtained for each query will be examined to determine if they are comparable, comparable will be further defined below. When examining the results from the two stoplists there are three possibilities that could arise concerning the number of documents, we might find:

1. Same number of results.
2. More results returned by the best stoplist.
3. More results returned by the general use stoplist.

A set of results will be categorized as the comparable if 1.) the results are exactly the same, meaning the same documents were returned, or 2.) the results which are different are topical with regard to the query which was submitted. Topical will be defined as being directly related to the subject of the specific query. As an example, if 16 documents were returned for a query when the general stoplist was being used, and only 10 were returned when the best stoplist was being used we would only consider the results comparable if the ten articles returned by the best stoplist were all topical with regard to the query used. That less documents were found is not as important as the documents which were returned are topical. The other possible situation would be if the best stoplist returned more documents than the general stoplist was used. Again we would need to determine if all the documents returned were topical. If every one of them was topical we would then conclude the results were comparable.

Each of the sets of queries will be evaluated using the method above and a total count of the number of comparable queries will be made. In the Results and Analysis section we will discuss the results of this analysis and make some observations concerning the number of comparable queries.

Summary

We have now outlined the method for creating our stoplist, and how they will be evaluated. In the next chapter we will build our stoplists and then determine which one of them is the most effective.

This chapter outlined the methods by which the research objectives of this thesis will be met. It describes the processes to be used to create a stoplist, and the statistical measures to be used to determine which one of the four stoplists best performed the functions of a stoplist. The following chapter will detail the creation of a stoplist, the testing of it and an in-depth analysis of the results obtained from testing. At the end, we will determine if the data gathered in the testing phase supports the hypotheses of this thesis.

IV. Results Analysis

Chapter Overview

In this chapter we will produce several stoplists as described in Chapter III. Next we will select the queries to use in testing the effectiveness of the new stoplists and then use these queries with the Air Force's homepage. Finally we will analyze the results of the queries and determine if we are able to support the hypotheses proposed in Chapter I.

Stoplist Creation

The following section will document the creation of four separate stoplists based on the steps outlined in the preceding chapter.

Corpus Selection

As stated in Chapter III, the literature reviewed does not outline a specific technique for selecting a corpus. We will use the guidelines from Chapter III to help choose which of the available corpuses we will use for our purposes.

General Usage Corpus. In Chapter III we established the following guidelines to use in selecting our general use corpus:

- Corpus should be larger than the Air Force corpus.
- Corpus should contain material from many areas.
- Format should be easy to work with.
- Cost of the corpus should be rather low.

I decided to use the British National Corpus (BNC) as the general use corpus because it was the only corpus found which met all the guidelines for selecting the general use corpus. The BNC contains 100 million unique words from over

4,000 documents, and represents works from nine areas. A frequency count of the unique words in the BNC was also available for public use in a format compatible with the available software. Information on the word frequency count of the BNC is available through the World Wide Web (Adam, 1996)

Air Force Usage Corpus. The guidelines we established for selecting the general corpus will also apply when selecting the Air Force corpus. The Air Force Electronic Publishing Library (AFEPL) was the corpus that best fit the criteria we established in Chapter III. The AFEPL is free, published approximately every three months, and contains documents from all the areas of operation within the Air Force. The AFEPL contains all the Policy Directives and implementation instructions for the USAF.

Determine Word Counts Because the word count for the BNC has already been determined it is imperative the rules and methods used to determine those counts be applied when determining the frequencies of the words in the Air Force Corpus.

BNC Rules. The word counts for the BNC were developed through analysis of the tagged version of the BNC (Adam, 1996). The major concern, when examining the BNC was how contractions were counted. Some corpus, or text analysis programs, truncate contractions and convert the word to its root. For example, the contraction "can't" would be listed as the root word "can". A search of the document revealed words such as "isn't," "can't" and "couldn't," although they appeared rather infrequently. The word "isn't" was found a total of ten times in the corpus giving it a frequency of 0.00000998941 percent. All the words counted were also recorded as lowercase regardless of their location within a sentence.

These rules are consistent with the way our word counting program, FREQUOUT, was written and no adjustments will be needed. We can now count the word in the Air Force corpus and compute the frequencies for each of the corpuses.

Word Counts. The Air Force corpus contained 52,241 unique words and consisted of 8,644,929 word occurrences. To use FREQOUT to count the words in the Air Force corpus those files stored as Microsoft Word documents were converted to text format and then processed. It was not necessary to count the words in the BNC since we obtained this information from another source but some modifications were necessary.

Modifications to BNC Word Counts. An examination of the words in the word count file revealed 21,237 entries were words combined with formatting information. It was apparent when viewing these strings of character what the embedded word was but the entries were still deleted. The entries were deleted because of the time it would have taken to correct the errors, and the entries accounted for 0.021 percent of the total words in the BNC.

Find Word Overlap

Having counted the number of times each of the words in both corpuses occurred we can now determine which words are common to both corpuses. As detailed in Chapter III, this will limit the number of words analyzed during the remaining steps of the process and eliminate acronyms and words spelled incorrectly.

Unfortunately, the time constraints of completing this thesis caused parts to be completed ahead of the documentation for other sections. The actual creation and testing of the stoplists we are now writing about have already been completed. This fact has lead to a problem with finding the word overlap between the two corpuses. The reasoning originally used to set-up the thesis failed to take into account one facet of this area. We will examine the proper way the overlap should have been determined. The remainder of the analysis section will be concerned with results obtained under the incorrect assumptions, and then at the end of this chapter we will examine what changes would have been made to each of the stoplists if the new criteria had been used.

Incorrect Assumptions. Earlier we assumed all the words that occurred in one corpus, but not in the other, would cause problems when calculating the measures of significance. The following paragraphs will discuss this problem with the unique words in each of the corpuses and establish what words will be included in further analysis.

Words Unique to the Air Force Corpus. Words appearing in the Air Force corpus and not in the BNC would cause problems when calculating two of the measures of significance. Measures of significance S2 and S4 would involve dividing by r, the frequency of the word in the BNC, when r would be equal to zero. For this reason it would be prudent to exclude words unique to the Air Force corpus from the calculation of S2 and S4. Figure 10 shows what the calculations of the measures of significance look like when considering these situations. Words that appear in the Air Force corpus, but not the BNC, should not be excluded from the calculation of S1 and S3. The calculations of S1 and S3 are changed somewhat for words unique to the Air Force corpus but it is still possible to calculate them.

When $r \neq 0$	When $r = 0$
$S1 = f - r$	$S1 = f$
$S2 = {}^{f}\!/_{r}$	Divide By zero error
$S3 = \frac{f}{f+r}$	$S3 = \frac{f}{f} = 1$
$S4 = \log\frac{f}{r}$	Divide By zero error

Figure 10. Adjusted Measures of Significance for Air Force Corpus unique words

Words Unique to the BNC. Words appearing in the BNC and not the Air Force corpus should not be part of further calculations since they are not used in the writing of Air Force documents. These words are not part of Air Force writing and they would most reasonably not make good stopwords since they would be considered very rare rather than very common. For this reason we will not include words unique to the BNC in further analysis.

Word Overlap Rules. In summary, the following guidelines would apply when determining the word overlaps in future exercises of this nature.

- Words unique to the BNC would be eliminated from analysis.
- Words unique to the Air Force corpus would be eliminated from the calculation of S2 and S4.

- Words unique to the Air Force corpus would be included in the calculation of S1 and S3.

For purposes of this thesis, the determination of which words were unique to each of the corpuses was done using Microsoft Access.

Calculate Word Frequencies

The word counts for each word in both corpuses was divided by the total number of words in the respective corpus. The BNC contained 99,894,105 words, while the Air Force corpus contained 8,644,929 words. For the words unique to the Air Force corpus the corresponding BNC frequency will be recorded as zero. The frequencies will be used to determine the objective of our analysis, the measures of significance.

Calculate Measures of Significance

The actual calculation of the measures of significance is rather simple and done by computer. Because of the number of words we are dealing with it is not possible or feasible to include the measures of significance for each of the words. The next section shows the graphs for each of the measures and provides some analysis of each graph.

Graph Measures of Significance

Our next step is to produce a graph, for each of the measures of significance, of S_i vs f_i for all words. The graphs will each have a different number of points determined by the overlap guidelines, discussed above. Figure 11 does not present clear and easy to understand results. Considering the examples shown in figure 6 it was reasonable to expect we would find a very clear and distinct pattern in each of the functions. It is very difficult to see any clear limits or cut-off points in any of the functions

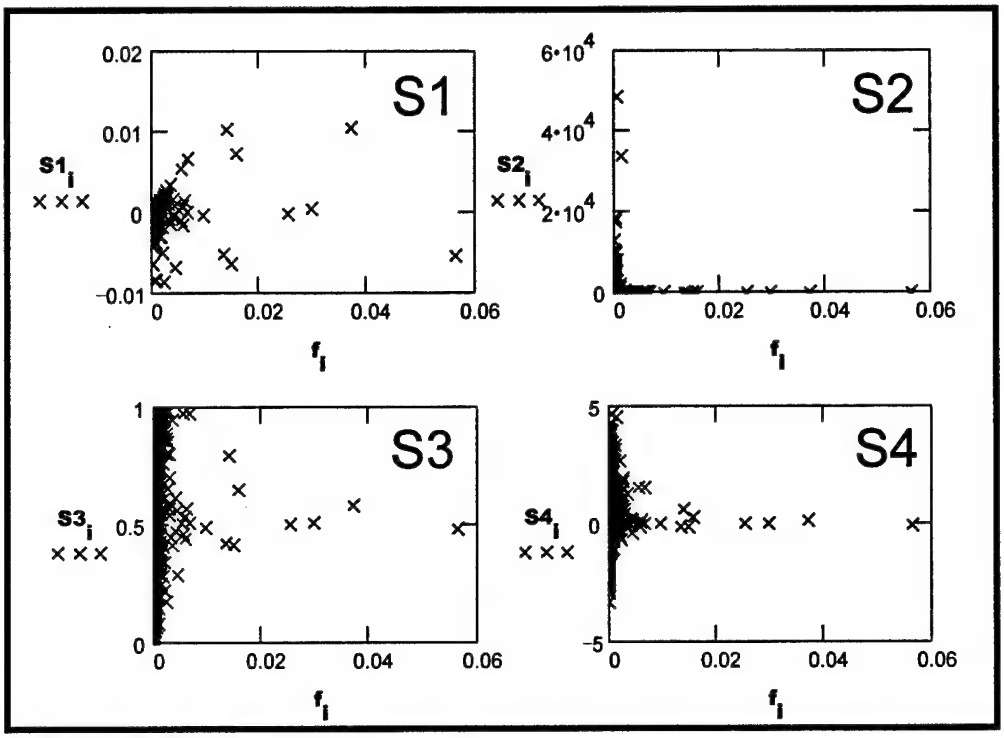


Figure 11. Measures of Significance for S1, S2, S3, and S4

Earlier in Chapter II we discussed the possibility of problems arising from our extending the theories upon which our approach is based. With relatively few words in an article, or single book, it would be quite easy to determine a cut-off point from each of the graphs. With the large number of point involved in this effort we are very hard pressed to determine any cut-off points.

Although we can not determine insignificance graphically, we can still use the measures of significance to create our stoplists. Our approach has been to compare the frequency with which a word occurs in one corpus to the frequency of its use in another corpus.

To help determine which measures are important to us we have created an example to show how this might be done. Table two below outlines the calculation of the measures of significance for three words. The words were chosen to represent various levels of comparative usage between BNC and Air Force corpus

Table 2. Sample Calculation of Measures of Significance

Word	Word Count		Frequency		Measures of Significance			
	BNC	AF Corpus	BNC	AF Corpus	S1	S2	S3	S4
compliment	74	4	0.0074	0.0040	-0.0034	0.5405	0.3509	-0.2672
the	1918	201	0.1918	0.2010	0.0092	1.0480	0.5117	0.0203
aircraft	10	200	0.0010	0.2000	0.1990	200.00	0.9950	2.3010
Total Number of words	10,000	1,000						

Looking at table 2 we see the measures of significance for "aircraft" are higher in each case when compared to the measures for each of the other words. In all cases, the measure for "compliment" was smaller than the measures for both "the" and "aircraft." This is a limited example but it is very safe to use larger values of the measures of significance to indicate insignificance. We can also see the computed value of AF frequency also matched the indications of insignificance we determined from the measure of significance. The higher the frequency for a word in the Air Force corpus the more insignificant it was in our initial description of the example.

We can now use the calculated measures of significance to indicate relative insignificance of a word. We will use words with greater measures of significance as our stopwords. It also appears there may be some value in using the frequency of a word from the Air Force corpus as an indication of insignificance

Limit Stoplist Size

In Chapter III we indicated our stoplists would be limited in size based on previous stoplists and the number of words we were examining after finding which words were common to both corpuses. Our stoplists will be limited to the first 425 words as ranked in descending order of their individual measures of significance The process of finding words common to both corpuses left us with 52,876 unique words. So it is reasonable to continue using a stoplist of 425 words. We are selecting a small portion of the total words under consideration and we are not making the stoplist larger than what the search program can accommodate. Figure 12 shows what our graphs of the measures of significance look for these 425 words. Limiting the number of words under consideration does not effect the appearance of these graphs. Limiting the stoplist to 425 terms allows us to examine the entire stoplists and compare them with each other for similarities. A very interesting fact is that measures of significance S2, S3, and S4 contain the same words. The words do not appear

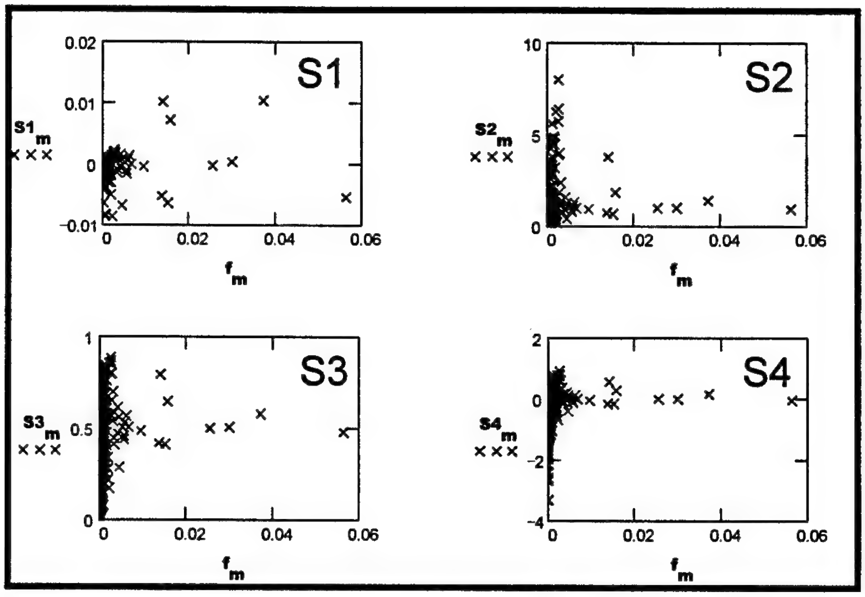


Figure 12. Plots of Measures of Significance for top 425 words in each stoplist

in the same order when listed according to their measure of significance, but they are the same. The implication for our effort is we now are operating with four unique stoplists. Our stoplists will now be identified as shown in table 3 below.

Table 3. Stoplist Names and Sources

Name	Source
BNC	British National Corpus
AF	Air Force Corpus
S1	Measure of Significance S1
S2S4	Measures of Significance S2, S3, and S4

Review Stoplists

Next, we will examine the stoplists and ensure we are not including words that are crucial to our application or if there are any words that are not present that should be.

Remove Significant Words.

The process of removing significant words is not scientific and relies heavily on familiarity with the subject area. Considering the subject area we are looking at only the word "command" seemed to be of great importance and reasonable to remove from all the stoplists. Command and all of its forms are essential to locating many articles of importance and identifying many people who might be mentioned within the Air Force homepage. Other words such as "tactical," "air," and "combat" were also considered as important but were determined to be used more often as verbs rather than proper names or a part of a title.

Add Significant Words.

The purpose of adding significant words is to ensure words that do occur frequently, even in general use, are included. The most obvious words looked for were "the," "and," "of," "a," and "and." all of which were present in all the stoplist. After removing the words indicated in the above paragraph the BNC stoplist contained 425 words, the AF stoplist contained 422 words, the S1 stoplist contained 421 words, and the S2S4 stoplist contained 425 words. The final stoplists are shown in Appendix B. Having completed all the steps to create our subject specific stoplist we are now ready to determine their effectiveness.

Queries

Our methodology stated all queries would be randomly selected from queries submitted by users to the Air Force homepage. Selection of the user submitted queries was done using the search logs from September 21,22, and 23 1997. The queries selected, as they were submitted to the search engine, are listed in Appendix C.

The other 30 queries were selected as described in Chapter III. These queries will be submitted enclosed in quotation marks since we are searching for the exact occurrence of the query not each of the individual words.

Summary

This chapter has focused on the creation of our stoplists and the selection of the queries we will use to test the effectiveness of our stoplist. Chapter V will discuss the results from the implementation of our testing strategy outlined in Chapter III.

V. Implementation and Analysis

Chapter Overview

In this chapter we will conduct the testing of the stoplists created in Chpater IV. We will submit each of the queries we selected in Chapter IV after the Air Force's homepage has been reindexed with each of the four stoplists we created. Finally we will analyze the results of the queries and determine if we are able to support the hypotheses proposed in Chapter I.

Data Gathering

The following paragraphs will document how the information to test effectiveness of each of the stoplists was obtained. Data gathering consisted of selecting the appropriate queries to use and recording the results returned when each one was submitted to the Air Force homepage search engine. The following paragraphs will outline how the results were recorded, and discusses any problems met during the process.

Number of Documents Returned

As each of the queries was submitted to the Air Force homepage search engine the results were recorded. Appendix D lists the number of documents returned for each of the queries submitted for each of the stoplists used. The number of documents returned indicates the actual number of documents found for each of the queries submitted. In some cases, the same document is cited twice within the same set of results. This is a function of the search engine and can not be corrected.

Information Returned

As the results from each of the queries were returned from the Air Force homepage search engine they were saved in a separate file and stored in a directory unique to the stoplist being used.

Search Engine and System Statistics

As outlined in Chapter III, the time to complete each query, the amount of central processor used, and the amount of memory used by each query was recorded. In those instances when a query occurred in more than one five second interval of the server statistics log, the information was taken from the latest interval in the log. These statistics are included in Appendix D.

Analysis of Results

The data necessary to evaluate the effectiveness of each of the stoplists we created has now been gathered. Using the methodology outlined in Chapter III we will examine each of the hypotheses and determine if the data gathered supports them.

Quantitative Analysis

Each of the three hypotheses of this thesis will be evaluated using the methodology outlined in Chapter III.

Hypothesis A Difference in Stoplist Contents

Research hypothesis A asserted that the words contained in stoplist generated from an Air Force corpus would be significantly different from a stoplist based on the most frequent words in a general usage corpus. Since we selected the BNC as our general usage corpus any comparison done to support or reject this hypothesis must be done using the BNC as the basis for comparison. For testing purposes our hypotheses are:

- Null Hypothesis: There is significant no difference between the words contained in a stoplist generated from the most common words in the BNC compared to stoplists generated from the Air Force corpus.

- Alternative Hypothesis: There is a significant difference in the words contained in stoplist generated from the BNC compared to stoplists generated from the Air Force corpus.

Table Four shows the number of words in common between each set of stoplists.

Table 4. Number of Words in Common between Stoplists

Overlap	AF	S1	S2S4
BNC	163	62	0
AF	-	318	15
S1	-	-	22

To either accept or fail to reject hypothesis A we should focus on the first row of the table. The first row of Table Four shows the largest number of words common between the BNC and any of the other stoplists is 163, or 38.4 percent. This would indicate we can reject our null hypothesis, and the words in a stoplist generated from an Air Force. From our analysis of hypothesis A we conclude we can reject the null hypothesis that a stoplist generated from an Air Force corpus is significantly different from a stoplist generated from a general usage corpus such as the BNC.

Hypothesis B Search Times

Hypothesis B is intended to show we can improve the speed of an application using a stoplist we have generated. We are trying to determine if a set of queries will be completed quicker when an AF corpus generated stoplist is used compared to the amount of time it takes to complete the same queries using the BNC stoplist. Unfortunately our data does not support rejection of the null hypothesis in this case.

The assertion of the hypothesis is the search times are better for our generated stoplists than the times from the BNC stoplist our null and alternate hypotheses are as follows:

- Null Hypothesis: There is no difference between the amount of time needed to complete a set of searches using the BNC stoplist and stoplists generated from the Air Force corpus.

- Alternative Hypothesis: It will take longer to complete a set of searches using the BNC than using stoplist generated from the Air Force corpus.

Table Five shows the results of the paired differences t-tests. From these results we fail to reject our null hypothesis and can not establish there is a difference in the amount of time needed to complete a set of query using the stoplist we have chosen.

Table 5. Paired differences t-tests between BNC and stoplists generated from AF corpus

	BNC-AF	BNC-S1	BNC-S2S4
Cases (n)	63	63	63
Test Statistic (T)	-1.8099	-1.5138	-1.45679
α	0.05	0.05	0.05
Rejection Region	t<-2.285 or t>2.285	t<-2.285 or t>2.285	t<-2.285 or t>2.285
Result	Fail to Reject	Fail to Reject	Fail to Reject

The data on the time to complete each of the searches is unique and requires some exploration and comment. After determining the amount of time to complete each of the queries it was apparent there was lack of variability in the data. Of the times used to conduct the t-test, a total of 252 queries, only ten of the queries took more than 5 seconds to complete. This seemed more than coincidental, especially considering the data that was eliminated earlier in the process. The manufacturer of the search engine used by the Air Force homepage was contacted to try determine a possible explanation. A manufacture's representative indicated the search engine log recorded the exact amount of time, to the nearest second, required to complete a search, and did not use a default value for any of the times recorded (Meyers, 1997). A log that recorded information with an accuracy greater than the nearest second would probably allow for some discrimination between the amount of time used by each of the stoplists.

Hypothesis C Number of Words Stoplisted from Queries

In Chapter II we highlighted the possible gains from reducing the sizes of indexes and queries by using a topic specific stoplist. This research hypothesis directly tests this by trying to establish which stoplist is significantly better at removing words from our randomly selected queries. Appendix E shows what the 80 test queries look like after they have had the stopwords removed.

Our null and alternate hypothesis are as follows:

- Null Hypothesis: BNC stoplist removes the same number of words from the same set of queries as the stoplists generated from the Air Force corpus.

- Alternative Hypothesis: BNC stoplist removes fewer words from the same set of queries than a stoplist generated from the Air Force corpus.

Our data allow us to reject the null hypothesis and assert our stoplist generated from the Air Force corpus removes significantly more words than the BNC stoplist. Table six outlines the results of our statistical analysis, and shows only AF and S1 removed significantly more words than the BNC stoplist.

Table 6. Results from paired differences t-tests between BNC stoplist and stoplists generated from AF corpus

	BNC-AF	BNC-S1	BNC-S2S4
Cases (n)	80	80	80
Test Statistic (T)	-6.24706	-0.776355	4.6788
α	0.05	0.05	0.05
Rejection Region	t<-1.99	t<-1.99	t<-1.99
Result	Reject	Fail to Reject	Reject

The next step is to see if there is a significant difference between the number of words removed by AF and S1.

Our null and alternate hypothesis are as follows:

- Null Hypothesis: S2S4 stoplist removes the same number of words from the same set of queries as the AF stoplist.
- Alternative Hypothesis: S2S4 stoplist does not remove the same number of words from the same set of queries as the AF stoplist.

This time we can see the AF stoplist removed a statistically significant larger number of words than the S1 stoplist did. Table Seven summarizes the results of the paired differences t-test for the comparison

Table 7. Results of paired differences t-test comparison of S2S4 and AF stoplists

	S2S4-AF
Cases (n)	80
Test Statistic (T)	-6.02094
α	0.05
Rejection Region	t<-2.2850 or t>2.2850
Result	Reject

Qualitative Analysis

Our last analytic task is to determine if a stoplist generated from the Air Force corpus is as effective as a general stoplist, such as our BNC stoplist.

Selection of Best Stoplist.

In Chapter III a method for selecting a best stoplist was outlined, and Table seven shows how each of the stoplists scored. From the results of our scoring we would conclude the AF stoplist was our best performer and will be used in our qualitative analysis section. We will now begin comparing the results returned for each query that meets the specifications outlined in Chapter III.

Table 8. Scores for best stoplist selection

	AF	S1	S2S4
Hypothesis B	20	20	20
Hypothesis C	10	30	20
Total	30	50	40

Comparison of Results Returned. Applying the guidelines in Chapter III limits the number of queries we are analyzing to just 37. In Chapter III we identified three categories that our results might fall into when comparing them. We will examine the results with respect to the three categories.

Same Number of Results. Those queries where each stoplist returned the same number of documents were put in this category. In these 33 cases the results would be classified as comparable. Each of the queries returned the same documents when both stoplists were used.

More results returned by the Best stoplist.

In only three cases were more documents returned by the Best stoplist than the BNC stoplist. Two of the queries returned results that were comparable to the results returned by the BNC stoplist and one did not. Queries 6, 49 and 57 all returned more documents when the Best stoplist was used. In queries 6 and 49 all the documents returned were topical to the query submitted and would be considered comparable results. In the case of query 57, the best stoplist returned 22 documents when the BNC stoplist only returned two documents. Only two of the documents returned by the best stoplist were topical to the original query, and the remaining results matched the words in the query but were not topical.

More Results Returned by the BNC Stoplist. One query returned more results from the BNC stoplist than from the Best stoplist. The analysis of the results would classify the documents returned as comparable. Query number 36 returned nine documents when the BNC stoplist was used and only eight documents when the Best stoplist was used. An examination of the results indicates some documents were listed more than once, resulting in an unequal number of documents being reported. When this is taken into account we find each stoplist returned the same results.

Analysis of Qualitative Results. The results of our qualitative analysis are listed in Table 8. Although not a conclusive finding, it would appear the majority of the queries returned comparable results when we compare the two stoplists. We have only examined a small portion of the queries we tested, but found the stoplist generated from the Air Force corpus produced the same results as the general use stoplist.

Table 9. Analysis of Number of Comparable Results by Category

	Number of Queries with Comparable Results	Number of Queries with Non-Comparable Results
Same number of results	33	0
More results returned by Best stoplists	2	1
More Results Returned by BNC stoplist	1	0
Total	36	1

Effects of Incorrect Assumptions

Earlier in this chapter we discovered an incorrect assumption was made in the early stages of the thesis. This error centered on an assumption about limiting our analysis to just the words common to the BNC and Air Force Corpus. Our earlier discussion concluded Air Force unique words would be included in the calculations of S1 and S3 but not in the calculations of S2 and S4. Measures of significance S1 and S3 were recomputed and the top 425 words were extracted and compared to the original words for both lists. The following paragraphs explain what differences were found in each of the stoplists.

Effects on S1

The exclusion of Air Force corpus unique words had a dramatic effect on the final composition of S1. The newer S1 only contained three of the same words contained in the original S1. A review of the words in the new S1 indicates it is a much more mathematical list, meaning it has no real reflection of the words used within the Air Force. Using the new S1 stoplist on our test queries removed 66 words compared to the original S1 that removed 165 words. Appendix F lists the word in the new S1 stoplist in alphabetic order. Our decision to exclude Air Force unique words seems an appropriate one that produced a more efficient stoplist.

Effects on S3

The effect on the determination of which words to include in S3 is also quite dramatic. Including words without a value for r automatically pushes all these words to a value of one. In our case there were now 14,196 words with a value of one, making it quite impossible to determine which words are most insignificant. The net result, concerning S3, is to highlight words that are unique to the AF corpus, a task that can be accomplished through easier means and nets no gain. Excluding the word unique to the Air Force corpus was again the correct decision, and led to having an S3 stoplist to use in our analysis.

Summary

We have now implemented the process we outlined and researched in the earlier portion of the thesis. Each of our hypotheses have now been tested with data from an actual implementation of each of the stoplists, and It is reasonable to say we have seen a difference in the performance of the stoplists generated from the Air Force corpus, compared to the performance of the stoplist generated from the general usage stoplist. The next chapter presents my conclusions and outlines some areas for future research.

VI. Conclusions and Recommendations

Introduction

The basic purpose of this thesis was to present an idea that might have some effect on our ability to deal with the massive amounts of information we are faced with today. Stoplists may play a very small role in what we do in our daily lives, but they can prove quite effective in the areas where they are used. We will discuss some areas that constrained the effectiveness of the research effort and what might be done to overcome them in future studies in this area. Next we will discuss each of the research objectives, whether or no they were met, and suggest some area for further research. Finally, we will provide some conclusions regarding the information presented and how it can best be used in the futures.

Scope and Limitations of the Study

The scope of this research effort was constrained by several factors beyond the control of those involved and may or may not be able to be surmounted in future efforts in this area.:

Limited Knowledge of Text Analysis. My limited understanding of text analysis and items such as collocation only allowed a very cursory application of the information contained in both of the corpuses. A deeper understanding of how to properly analyze and create a corpus would allow information that was beyond me grasp to be used to produce a better corpus.

Availability of a Suitable Corpus of Air Force Language. The Air Force's Electronic Publishing Library was suitable for our purposes, but a more general collection of documents may have provided a more accurate representation of the way in which people in the Air Force write. Since the target application for a stoplist is the best source of a corpus, it might have been better if we could have performed our word count on all the documents accessible through the Air Force homepage.

Accuracy of the Search Engine Log. The search engine used by the Air Force homepage did not collect information in manner conducive to statistical analysis. The lack of accuracy beyond a second produced results that seemed to indicate the majority of all queries performed were executed in the same amount of time. When this information is examined closely it becomes clear a more accurate means of determining the time it takes to complete a query could provide information that could lead to different conclusions in some areas.

Research Objective 1

> Locate and summarize the documented methods for creating a stoplist from a corpus.

Chapter II summarized a diligent and thorough scouring of the available research assets yielded little to no published information on the creation of stoplists. The basic methodology for creating a stoplist does not seem to be of interest to those people in the field of information analysis and has been ignored. The process for creating a stoplist outlined by Fox (1989) was the only complete summarization found and was adopted as the model for our process.

Research Objective 2

> Develop and test a simple (2-3 step) process for creating a stoplist tailored to a subject area.

Chapter III outlined the process we were going to use for our research, and provided the reasoning for each of the steps taken. After reading the steps taken to produce a stoplist most people would probably begin to question the definition of the word "simple." The implementation of the process and testing of the resultant stoplists may also seem more difficult than expected. Unfortunately, working with a large amount of data requires some unique solutions, but these are needed given the software available to computer users today. A more appropriate means of creating a stoplist was also discovered and tested; a discovery that seemed to provide the most effective stoplist.

Research Objective 3

> Determine the effectiveness of a corpus specific stoplist through statistical comparison.

In Chapter IV we also tested the results of implementing each of our stoplists using the methods outlined in Chapter III. Although only two of our tests of hypotheses yielded conclusions we felt were significant we felt safe in saying one of our stoplists was more efficient than the other using the test queries we selected. Since this is truly an investigative study we can not draw conclusions which are substantial. I believe we can say there is reason to continue research into the effectiveness of subject area corpuses.

Recommendations

With the limited amount of research done on the creation and implementation of subject area stoplists it would be rather easy to not recommend areas for further research. However, I believe research in any of the following area may help prove the need for such stoplists, and the benefits or their use.

- Quantification of the actual cost savings from using a corpus specific stoplist. Today it is very difficult to obtain the support of new technologies or methods unless a return on investment can be demonstrated. An in-depth study of the financial costs of searching and indexing would allow a quantification of the benefits from using a corpus specific stoplist with an application
- Implement the Stoplist Creation Process Using a Narrower Application Although the Air Force homepage is very limited in its scope, dealing strictly with Air Force subjects, choosing an application with a narrower focus may provide more interesting results. An example of such an application would be an intelligence database, a medical treatment database, or even an electronic document storage system. Choosing this type of application, one that would also have a more narrow corpus, might produce results more in line with those graphed in figure 1.

Summary

Our efforts have shown there are possible gains to be made in many applications by reducing the amount of time it takes to complete a search. Our future seems to hold the promise of more information, but to make use of this information we need to access it faster and obtain what we want sooner. The amount of time saved by using a corpus specific stoplist may be quite small, but will add up to a significant amount of time over the long haul. Anything we can do to help computer users get what they want faster will be greatly appreciated by everyone in the information age.

Appendix A: FREQOUT Perl Script

The PERL script below was used to create the word count from the documents in the Air Force corpus. The files were in either in ASCII format or SGML.

```
#!/usr/local/bin/perl

%freq = ();
$word_count = "";

if ($ARGV[0] =~ /-f/) {
        shift;
        $file = shift;
        open(FREQ, "<$file");
        %freq = map {split} <FREQ>;
        close FREQ;
        delete $freq{"Word-count"};
}

#for (keys %freq) {
#       print "$_ $freq{$_}\n";
#}
#exit 0;

traverse_dir($ARGV[0]);
$/ = '';
for $key (sort keys %freq) {
        $word_count += $freq{$key};
        print "$key $freq{$key}\n";
}
print "Word-count $word_count\n";

sub traverse_dir {
        local($dir) = @_;
        local(@entries,$entry,);
        chdir $dir || die("cannot chdir");
        opendir(DIR, ".");
        @entries = readdir(DIR);
        for $entry (@entries) {
                next if ($entry =~ /^\./);
                next unless $entry;
                if (-d $entry) {
                        traverse_dir($entry);
                        chdir "..";
                }
                elsif ($entry =~ /^.*\.(txt|sgm)$/i) {
                        do_freq_count($entry);
                }
        }
        closedir DIR;
}

sub do_freq_count {
        local($file) = @_;
        $/ = '';

        open(IN, $file) || die "cannot open $file";
        while (<IN>) {
                s/<[^\r]*>\r//g;
                tr/A-Z/a-z/;
                @words = split(/\W*\s+\W*/, $_);
                for $word (@words) {
                        next if (length $word > 19);
                        next unless $word =~ (/^[a-z]+$/);
                        $freq{$word}++;
                }
        }
        close IN;
}
```

Appendix B. Stoplists

BNC

a
able
about
above
across
act
action
actually
after
again
against
age
all
almost
already
also
although
always
am
among
an
and
another
any
anything
are
area
areas
around
as
asked
at
available
away
back
be
became
because

AF

a
about
above
access
according
acquisition
act
action
actions
active
activities
activity
additional
administrative
af
afb
afi
afman
afr
after
agencies
agency
air
aircraft
all
also
an
analysis
and
any
applicable
application
apply
approach
appropriate
approval
approved
are

S1

access
according
accounting
acquisition
action
actions
active
activities
activity
additional
administration
administrative
aerospace
af
afb
afi
afman
afr
agencies
agency
air
aircraft
all
an
analysis
and
ang
annual
any
applicable
application
apply
appropriate
approval
approved
are
area
areas

S2S4

aan
accordance
according
accouterments
acronyms
acsc
advisement
aecs
aeromedical
af
afb
afcs
afftc
afi
afis
afm
afman
afms
afo
afr
afs
agard
aia
aiguillette
airspeeds
alc
alcm
alcms
alcs
amc
amn
analyzes
ang
annotates
annuitant
annunciator
ao
apf

become
been
before
began
behind
being
best
better
between
big
black
body
book
both
britain
british
business
but
by
ca
called
came
can
car
care
case
centre
change
child
children
city
clear
come
community
company
control
could
council
country
court
day
days
development

area
areas
as
assigned
assignment
assistance
at
attachment
auth
authority
authorized
available
award
b
base
based
basic
be
because
been
before
being
between
board
both
budget
but
by
c
can
care
career
case
category
center
change
changes
chapter
chief
civil
civilian
class
code

as
assigned
assignment
assistance
associated
attachment
auth
authority
authorized
available
award
b
base
based
basic
block
board
budget
c
cannot
capabilities
capability
career
category
center
certification
change
changes
chapter
check
chief
civil
civilian
class
classification
classified
clearance
code
combat
communications
complete
completed
completion

asd
asr
atch
atd
attendee
attn
auth
authenticators
authorizations
bcas
bce
behaviors
berm
berms
biss
bldg
blvd
bolling
bpa
bpas
brf
caliber
calibrates
canceled
cannot
catex
cbl
ccts
cem
centers
centimeters
cercla
certifications
cmcs
cmip
coa
cognizant
collocated
compensable
comsec
concessionaire
conus
copilot

did
different
difficult
do
does
doing
done
door
down
during
each
early
economic
education
effect
either
end
england
english
enough
er
even
ever
every
eyes
face
fact
family
far
father
feel
felt
find
first
five
following
for
form
found
four
from
full
further

combat
communications
complete
completion
computer
conditions
conduct
construction
contract
control
copy
cost
costs
course
courses
criteria
cs
current
d
data
date
day
days
dd
defense
department
design
designated
destroy
determine
develop
development
direct
do
document
documents
dod
does
during
duties
duty
e
each

compliance
component
components
computer
conditions
conduct
construction
contact
contingency
contract
contracting
contractor
control
coordinates
coordination
copies
copy
cost
costs
course
courses
criteria
cs
current
d
data
date
days
dd
defense
department
design
designated
destroy
determine
develop
development
develops
direct
directive
directives
discharge
distribution

counseling
counselor
counterterrorism
courseware
cpf
crating
crewmembers
crm
crosscheck
cryogenics
crypto
cryptographic
csra
csso
ctf
ctn
cto
cwf
daa
dd
declassify
defense
defenses
deployable
designator
designees
disa
dishonorable
dla
dod
dosimeter
dru
dsn
dtg
eals
eci
ecl
ecos
eeo
ef
eglin
englewood
enlistment

future
gave
general
get
give
given
go
god
going
good
got
government
great
group
had
half
hand
hard
has
have
having
he
head
health
held
help
her
here
high
him
himself
his
home
house
how
however
i
if
important
in
including
information
interest

education
ef
effective
emergency
emphasis
employee
employees
engineering
ensure
enter
entry
environmental
equipment
established
etc
evaluation
example
experience
f
facilities
facility
federal
field
figure
file
final
financial
first
flight
flying
following
for
force
forces
foreign
form
format
forms
from
functional
functions
fund
funds

division
document
documentation
documents
dod
during
duties
duty
e
each
education
ef
effective
electronic
eligible
emergency
emphasis
employee
employees
engineer
engineering
enlisted
ensure
ensures
enter
entry
environmental
equipment
establish
established
establishes
etc
evaluation
example
experience
f
facilities
facility
federal
field
figure
file
files

enlistments
enrollment
eod
eom
eoq
epaulets
eql
equivalency
escp
essentiality
esv
ew
examinee
extemporaneous
faf
familiarization
familiarizes
fams
fars
favorable
fcf
fdt
fers
fh
flightline
flir
flowsheet
fma
foa
followup
fpm
fpo
fso
func
fwa
fwg
fws
fy
gcm
geodesy
gfe
glac
glideslope

international
into
is
it
its
itself
job
john
just
keep
kind
knew
know
known
large
last
later
law
left
less
let
level
life
like
likely
line
little
local
london
long
look
looked
looking
love
made
main
major
make
making
man
many
market
may

general
government
grade
ground
group
guidance
has
have
he
health
her
his
hours
how
however
hq
i
identify
if
in
include
includes
including
individual
individuals
information
initial
installation
instruction
instructions
intelligence
into
is
issues
it
item
items
its
job
joint
knowledge
least
leave

financial
flight
flying
following
for
force
forces
foreign
form
format
formerly
forms
functional
functions
fund
funds
general
grade
group
guard
guidance
hazardous
headquarters
health
host
hours
hq
identification
identified
identify
if
include
includes
including
individual
individuals
information
initial
inspection
installation
installations
instruction
instructional

gsa
gtr
handtools
hazmat
heliports
hics
honorable
honors
hpp
hq
humint
hvac
icbm
ids
idt
ima
immunizations
indorsement
indorses
instructional
instrumentalities
interns
intl
irm
irvington
isd
ivd
jcs
jka
jprs
jss
kirtland
kristy
lackland
lccs
leadtime
lethality
lft
lmr
loadmaster
loas
lof
logisticians

me
mean
means
members
men
might
million
mind
minister
mm
moment
money
months
more
most
mother
mr
much
must
my
name
national
need
never
new
next
night
no
not
nothing
now
number
of
off
office
often
oh
old
on
once
one
only
open

level
line
list
local
location
logistics
made
maintain
maintenance
major
make
management
manager
manpower
manual
material
materials
may
medical
meet
member
members
message
military
minimum
mission
months
more
most
must
naf
name
national
necessary
need
needed
needs
new
no
normally
not
note
nuclear

instructions
instructor
intelligence
inventory
issues
item
items
joint
knowledge
l
least
level
limited
list
listed
location
logistics
maintain
maintains
maintenance
major
management
manager
managers
mandatory
manpower
manual
material
materials
materiel
may
medical
meet
member
members
message
military
minimum
mission
must
naf
name
national

longline
lors
lox
lsc
maca
mailman
maintainability
mandays
maneuvers
maso
materiel
mcm
meep
metar
mfc
millimeters
minimums
misbehavior
mns
moa
modeling
mopp
mou
mre
mso
mtl
mva
mwr
mws
naci
naf
nafi
nafis
navaid
navaids
ncoic
nfpa
noa
noncommissioned
noncompetitive
notam
notams
notifies

or
order
other
others
our
out
over
own
part
particular
particularly
party
past
pay
people
perhaps
period
person
place
point
police
policy
political
position
possible
power
present
probably
problem
problems
process
provide
public
put
question
quite
read
real
really
report
research
result
right

number
objectives
of
office
officer
officers
official
on
one
only
operating
operation
operational
operations
or
order
orders
organization
other
out
over
paragraph
part
pay
per
perform
performance
period
person
personal
personnel
physical
place
plan
planning
plans
point
policies
policy
position
positions
possible
power

necessary
needed
needs
normally
note
notification
nuclear
number
objective
objectives
of
office
officer
officers
official
operating
operation
operational
operations
opr
or
order
orders
organization
organizations
other
overseas
paragraph
pay
per
perform
performance
personnel
phase
physical
plan
planning
plans
policies
policy
position
positions
prepare

nrb
nvg
obligating
odor
oes
offense
offenses
ois
opd
opm
opr
ords
osc
osha
ota
overgarment
overpressure
oversea
paa
paq
paralegal
paralegals
pardini
pdo
pensacola
pertain
pertaining
pfe
pgm
physicals
pid
pmd
pme
pmf
poc
poi
pov
ppbs
preflight
prepositioning
preproduction
prerequisites
pretest

road
room
round
run
said
same
saw
say
says
school
second
see
seemed
seen
sense
service
services
set
several
she
should
show
side
since
six
small
so
social
society
some
something
south
special
staff
start
state
still
study
such
support
sure
system
take

primary
prior
procedures
process
processing
program
programs
project
property
protection
provide
provided
provides
public
publications
purpose
quality
record
records
ref
related
report
reporting
reports
request
requests
require
required
requirements
reserve
resources
responsibilities
responsibility
responsible
review
safety
same
school
secretary
section
security
see
service

prepares
prerequisites
primary
prior
procedures
process
processing
program
programs
project
promotion
proper
property
protection
provide
provided
provides
publications
purpose
qualification
qualified
quality
quota
readiness
receive
record
records
ref
refer
related
repair
report
reporting
reports
request
requests
require
required
requirement
requirements
requires
reserve
resource

pursuant
radiologic
rater
raters
rcs
rdd
reassignment
reassignments
recertification
recordkeeping
recoupment
reemployment
reengineering
reentry
registrant
registrants
reimbursable
reimbursements
rekeying
releasable
relocatable
remotivation
reorganizes
reparable
reportable
reprogramming
requester
requesters
retitle
revalidate
revetment
revetments
rsd
rso
sabc
saf
samis
samm
sanitization
saos
sapm
sav
sbi

taken
taking
tell
than
that
the
their
them
themselves
then
there
therefore
these
they
thing
things
think
this
those
though
thought
three
through
time
times
to
today
together
told
too
took
top
towards
turned
two
under
until
up
upon
us
use
used
using

services
she
should
skills
so
some
space
special
specialty
specific
staff
standard
standards
state
statement
states
status
storage
student
students
subject
such
supply
support
system
systems
table
team
technical
test
than
that
the
their
them
then
there
these
they
this
those
through
time

resources
responsibilities
responsibility
responsible
review
reviews
safety
secretary
section
security
selected
selection
send
senior
separation
serves
service
services
should
skills
software
space
special
specialty
specific
specified
squadron
staff
standard
standards
statement
states
status
storage
student
students
submit
such
summary
supervisor
supply
support
supporting

sbp
sct
sdc
sdt
sealift
sei
seis
selectees
semiannual
semiannually
serviceability
sja
slbms
soa
sof
speci
specialty
splinting
sptc
srp
srr
ssf
ssn
sso
sss
stepparent
sts
stt
subparagraph
subparagraphs
subproject
supportability
survivability
sustainment
sv
takeoffs
tasked
tasking
taxiways
tdy
tfa
theater
tlf

very
view
voice
want
wanted
war
was
water
way
we
week
well
went
were
what
when
where
whether
which
while
white
who
whole
why
will
with
within
without
woman
women
words
work
working
world
would
yeah
year
years
yes
yet
you
young
your

title
to
total
training
transportation
travel
two
type
under
unit
united
units
up
upon
us
usaf
use
used
using
vehicle
was
water
we
weapons
what
when
where
which
who
will
with
within
without
work
would
x
year
years
you
your

system
systems
table
task
tasks
tdy
team
technical
test
testing
these
through
title
total
training
transfer
transportation
travel
type
u
under
unit
united
units
usaf
use
used
using
vehicle
vehicles
volume
waste
weapon
weapons
weather
when
will
within
x

tmo
tna
tos
tpr
tpt
traveler
travelers
traveling
trc
troubleshooting
troubleshoots
trs
tryout
tryouts
tts
tx
uln
umd
unfavorable
unsuspended
upt
usaf
usafa
usafe
usps
usss
utc
uucp
vapor
vapors
vectoring
verifies
vk
vortac
waivers
wargaming
wastewater
weaponeering
wideband
willful
wk
workdays
wps

Appendix C. Search Queries

Queries 1-50 were randomly selected from quereis entered through the Air Force homepage Search functions. As described in Chapter III, queries 51-80 were randomly selected from items located in various sections of the Air Force homepage. Queries 1-50 were altered to correct for spelling errors, and may have had quotation marks or BOOLEAN functions added, such as AND and OR, or search operators to make the query search for more specific items

1 SPA147
2 scouts
3 "Samuel C. Robbins"
4 ipms
5 air ADJ logistics
6 AMC AND flight AND dependent
7 "headquarters USAF"
8 earned
9 "Joint Staff"
10 "reconnaissance aircraft"
11 "early separation"
12 "Ellins Airbase"
13 "Retired Pay" AND offset
14 "Office of primary responsibility"
15 b36 AND peacemaker
16 "kenneth lowry"
17 fox AND edwardAND j AND 15471683
18 tuition AND assistance AND changes
19 reqroting
20 mechanics
21 "keflavik, iceland" AND ab
22 "Juan A. Marrero"
23 winemiller
24 "damage control"
25 VC-25A
26 "E-Mail Listings"
27 "Air Force" AND artwork AND emblem AND logo
28 "Michael Egan"
29 "Special Tactics Team"
30 "ft. hood"
31 "earnest harmon air force base"
32 stealth
33 "Space Maneuver Vehicle"
34 Elmendorf
35 "tandem thrust"
36 accident AND shaw
37 "Jackie Parker"
38 "military paychart"
39 "www.afmc.wpafb.af.mil/lean/nsn ."
40 "10th Tactical reconnaissance Wing"
41 "General Officers"
42 closing AND lowry
43 "military retired pay center"
44 Lompoc
45 "afi 33-270"
46 412
47 Jackson AND Charles
48 elliott
49 50th AND Air AND Force AND Logo
50 air AND force AND instruction AND 32-7064

51 "Presidential Maintenance Branch"

52 "air traffic control for establishing air assault landing zones, close air support for strike aircraft "

53 "launched into geosynchronous orbit on a Titan IV booster"

54 "Air Force investigated Unidentified Flying Objects"

55 "vice commander, 22nd Air Force"

56 "the operation of the USAF Test Pilot School"

57 "1991 Defense Systems Management College"

58 "aide to the commandant Air War College"

59 "vice commander Headquarters U.S. Air Forces in Europe"

60 "25th Air Division, Tactical Air Command, McChord"

61 "Keeping the peace in Sarajevo"

62 "Pegasus launches experimental payload"

63 "Tanker units rotate in south of France"

64 "Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994"

65 "The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land"

66 "Top civil servants to enter education, development program"

67 "Air Force is a leading-edge customer"

68 "General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators"

69 "obtain military memorabilia such as patches, uniform items, mugs"

70 "Navy Seals egress from a MH-53J Pave Low in the mountains of Norway"

71 "Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2"

72 "Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)"

73 "Chief of Staff visits troops in Tuzla"

74 "The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training."

75 "any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations"

76 "performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world"

77 "It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display"

78 "The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces."

79 "General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010"

80 "If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices"

Appendix D: Query Statistics

Number of Hits

Query	BNC	AF	S1	S2S4
1 SPA147	0	0	0	0
2 scouts	27	27	27	27
3 "Samuel C. Robbins"	0	0	0	0
4 ipms	0	0	0	0
5 air ADJ logistics	165	0	0	164
6 AMC AND flight AND dependent	3	4	4	25
7 "headquarters USAF"	5	-	0	-
8 earned	-	-	-	-
9 "Joint Staff"	-	0	0	146
10 "reconnaissance aircraft"	52	-	-	52
11 "early separation"	147	1	-	1
12 "Ellins Airbase"	0	0	0	0
13 "Retired Pay" AND offset	2	2	2	0
14 "Office of primary responsibility"	13	0	0	5
15 b36 AND peacemaker	0	1	1	0
16 "kenneth lowry"	0	0	0	0
17 fox AND edwardAND j AND 15471683	0	0	0	0
18 tuition AND assistance AND changes	6	67	67	6
19 reqroting	0	0	0	0
20 mechanics	59	59	20	59
21 "keflavik, iceland" AND ab	6	6	21	6
22 "Juan A. Marrero"	0	0	0	0
23 winemiller	0	0	0	0
24 "damage control"	-	-	-	0
25 VC-25A	10	10	4	4
26 "E-Mail Listings"	1	1	1	1
27 "Air Force" AND artwork AND emblem AND logo	0	0	0	0
28 "Michael Egan"	0	0	0	0
29 "Special Tactics Team"	5	-	-	5
30 "ft. hood"	4	4	4	4
31 "earnest harmon air force base"	0	0	0	0
32 stealth	-	-	-	-
33 "Space Maneuver Vehicle"	4	120	120	4
34 Elmendorf	192	192	193	193
35 "tandem thrust"	4	4	4	4

36	accident AND shaw	9	8	8	8
37	"Jackie Parker"	1	1	1	1
38	"military paychart"	0	0	0	0
39	"www.afmc.wpafb.af.mil/lean/nsn."	0	0	0	0
40	"10th Tactical reconnaissance Wing"	3	3	3	3
41	"General Officers"	-	0	0	46
42	closing AND lowry	3	3	3	3
43	"military retired pay center"	0	-	-	0
44	Lompoc	4	4	4	4
45	"afi 33-270"	0	0	0	0
46	412	32	33	32	32
47	Jackson AND Charles	14	14	14	14
48	elliott	2	2	2	2
49	50th AND Air AND Force AND Logo	8	9	9	9
50	air AND force AND instruction AND 32-7064	0	0	0	0
51	"Presidential Maintenance Branch"	1	1	1	1
52	"air traffic control for establishing air assault landing zones, close air support for strike aircraft "	1	1	1	1
53	"launched into geosynchronous orbit on a Titan IV booster"	1	1	1	1
54	"Air Force investigated Unidentified Flying Objects"	1	1	1	1
55	"vice commander, 22nd Air Force"	1	1	1	1
56	"the operation of the USAF Test Pilot School"	1	-	1	1
57	"1991 Defense Systems Management College"	2	22	22	2
58	"aide to the commandant Air War College"	0	0	0	0
59	"vice commander Headquarters U.S. Air Forces in Europe"	0	0	3	0
60	"25th Air Division, Tactical Air Command, McChord"	1	1	1	1
61	"Keeping the peace in Sarajevo"	0	0	0	0
62	"Pegasus launches experimental payload"	2	2	2	2
63	"Tanker units rotate in south of France"	2	2	2	2
64	"Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994"	*	*	*	*

65 "The nose landing gear of an EC-135 command and control aircraft collapsed w1hile attempting to land"	1	1	1	1
66 "Top civil servants to enter education, development program"	2	2	1	2
67 "Air Force is a leading-edge customer"	2	2	2	2
68 "General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators"	2	2	2	2
69 "obtain military memorabilia such as patches, uniform items, mugs"	-	0	0	0
70 "Navy Seals egress from a MH-53J Pave Low in the mountains of Norway"	1	1	1	1
71 "Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2"	1	1	1	1
72 "Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)"	2	2	2	2
73 "Chief of Staff visits troops in Tuzla"	2	2	2	2
74 "The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training."	1	1	1	1
75 "any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations"	*	*	*	*
76 "performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world"	2	2	2	2
77 "It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display"	1	1	1	1

78	"The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces."	*	*	*	*
79	"General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010"	-	0	0	0
80	"If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices"	5	5	5	2

Time

Query	BNC	AF	S1	S2S4
1 SPA147	5	5	5	5
2 scouts	5	5	5	5
3 "Samuel C. Robbins"	5	5	5	5
4 ipms	5	5	5	5
5 air ADJ logistics	6	5	5	8
6 AMC AND flight AND dependent	5	41	5	5
7 "headquarters USAF"	5	-	5	-
8 earned	-	-	-	-
9 "Joint Staff"	-	5	5	43
10 "reconnaissance aircraft"	6	-	-	9
11 "early separation"	5	5	52	5
12 "Ellins Airbase"	5	5	5	5
13 "Retired Pay" AND offset	5	5	5	5
14 "Office of primary responsibility"	5	5	5	5
15 b36 AND peacemaker	5	5	5	5
16 "kenneth lowry"	5	5	5	5
17 fox AND edwardAND j AND 15471683	5	5	5	5
18 tuition AND assistance AND changes	5	7	10	5
19 reqroting	5	5	5	5
20 mechanics	6	12	6	15
21 "keflavik, iceland" AND ab	5	5	5	5
22 "Juan A. Marrero"	6	5	5	5
23 winemiller	*	5	5	5
24 "damage control"	-	-	-	5
25 VC-25A	5	5	5	5
26 "E-Mail Listings"	5	5	5	5
27 "Air Force" AND artwork AND emblem AND logo	5	5	5	5
28 "Michael Egan"	5	5	5	5
29 "Special Tactics Team"	5	-	-	5
30 "ft. hood"	5	5	5	5
31 "earnest harmon air force base"	5	5	5	5
32 stealth	*	-	-	-
33 "Space Maneuver Vehicle"	5	25	35	5
34 Elmendorf	6	27	44	23
35 "tandem thrust"	5	5	5	5
36 accident AND shaw	5	5	5	5
37 "Jackie Parker"	5	5	5	5
38 "military paychart"	5	5	5	5

39 "www.afmc.wpafb.af.mil/lean/nsn."	5	5	5	5
40 "10th Tactical reconnaissance Wing"	5	5	5	5
41 "General Officers"	*	5	5	7
42 closing AND lowry	5	5	5	5
43 "military retired pay center"	5	-	-	5
44 Lompoc	5	5	5	5
45 "afi 33-270"	5	5	5	5
46 412	5	5	5	5
47 Jackson AND Charles	5	5	5	5
48 elliott	5	5	5	5
49 50th AND Air AND Force AND Logo	5	5	5	5
50 air AND force AND instruction AND 32-7064	5	5	5	5
51 "Presidential Maintenance Branch"	5	5	5	5
52 "air traffic control for establishing air assault landing zones, close air support for strike aircraft "	5	5	5	5
53 "launched into geosynchronous orbit on a Titan IV booster"	5	5	5	5
54 "Air Force investigated Unidentified Flying Objects"	5	5	5	5
55 "vice commander, 22nd Air Force"	5	5	5	5
56 "the operation of the USAF Test Pilot School"	5	5	5	5
57 "1991 Defense Systems Management College"	5	5	6	5
58 "aide to the commandant Air War College"	5	5	5	5
59 "vice commander Headquarters U.S. Air Forces in Europe"	5	5	5	5
60 "25th Air Division, Tactical Air Command, McChord"	5	5	5	5
61 "Keeping the peace in Sarajevo"	5	5	5	5
62 "Pegasus launches experimental payload"	5	5	5	5
63 "Tanker units rotate in south of France"	5	5	6	5
64 "Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994"	*	*	*	*
65 "The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land"	5	5	5	5

66	"Top civil servants to enter education, development program"	5	5	5	5
67	"Air Force is a leading-edge customer"	5	5	5	5
68	"General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators"	5	5	5	5
69	"obtain military memorabilia such as patches, uniform items, mugs"	-	5	5	5
70	"Navy Seals egress from a MH-53J Pave Low in the mountains of Norway"	5	5	5	5
71	"Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2"	5	5	5	5
72	"Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)"	5	5	5	6
73	"Chief of Staff visits troops in Tuzla"	5	5	5	5
74	"The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training."	5	5	5	5
75	"any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations"	*	*	*	*
76	"performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world"	5	5	5	5
77	"It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display"	5	5	5	5
78	"The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces."	*	*	*	*

79	"General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010"	-	5	5	5
80	"If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices"	5	5	5	5

CPU Usage

	Query	BNC	AF	S1	S2S4
1	SPA147	0.1	0.3	0.3	*
2	scouts	0.1	0.1	0.8	*
3	"Samuel C. Robbins"	0.1	0.3	0.1	*
4	ipms	0.3	0.1	0.1	*
5	air ADJ logistics	0	0.1	0.1	3.4
6	AMC AND flight AND dependent	0	0.1	0.1	*
7	"headquarters USAF"	0.1	-	0.3	-
8	earned	*	-	-	-
9	"Joint Staff"	*	0.3	4.2	2.7
10	"reconnaissance aircraft"	1.3	-	-	1.1
11	"early separation"	2.7	0.1	-	0.3
12	"Ellins Airbase"	0.1	0.1	3.1	*
13	"Retired Pay" AND offset	0.1	0.1	0.1	0.3
14	"Office of primary responsibility"	0.1	0.1	0.1	*
15	b36 AND peacemaker	0.3	0.4	0.1	*
16	"kenneth lowry"	0.1	0.1	0.1	*
17	fox AND edwardAND j AND 15471683	0.3	0.1	0.1	0.3
18	tuition AND assistance AND changes	*	1.3	0.3	0.2
19	reqroting	0.1	0.1	1.4	0.1
20	mechanics	1.3	1	0.1	0.8
21	"keflavik, iceland" AND ab	0.1	0.4	0.1	*
22	"Juan A. Marrero"	*	0.1	0.4	0.1
23	winemiller	*	0.1	0.1	0.2
24	"damage control"	-	-	*	0.4
25	VC-25A	0.1	0.1	4.2	*
26	"E-Mail Listings"	0.1	0.1	0.5	*
27	"Air Force" AND artwork AND emblem AND logo	0.4	0.3	0.4	*
28	"Michael Egan"	0.1	*	0.1	*
29	"Special Tactics Team"	0.1	-	-	*
30	"ft. hood"	0.1	0.1	4.2	0.3
31	"earnest harmon air force base"	0.3	0.3	0.1	0.3
32	stealth	*	-	-	-
33	"Space Maneuver Vehicle"	0.4	2.4	3.7	*
34	Elmendorf	3.3	3.9	2.5	3.9
35	"tandem thrust"	0.4	0.3	3.5	0.1
36	accident AND shaw	0.1	0.1	0.1	0.4
37	"Jackie Parker"	0.3	0.1	0.5	0.4
38	"military paychart"	0.1	0.1	0.2	*

39	"www.afmc.wpafb.af.mil/lean/nsn."	0.1	0.3	0.1	*
40	"10th Tactical reconnaissance Wing"	0.3	0.1	0.3	0.4
41	"General Officers"	*	0.3	0.2	1
42	closing AND lowry	0.1	0.1	0.1	*
43	"military retired pay center"	0.3	-	-	*
44	Lompoc	0.1	*	4.5	*
45	"afi 33-270"	0.1	0.1	0.1	0.3
46	412	0.1	0.1	0.1	0.8
47	Jackson AND Charles	0.1	0.1	0.9	0.1
48	elliott	0.3	0.4	0.1	*
49	50th AND Air AND Force AND Logo	0.6	0.1	0.1	*
50	air AND force AND instruction AND 32-7064	0.3	0.1	0.1	0.1
51	"Presidential Maintenance Branch"	0.3	0.4	0.1	0.4
52	"air traffic control for establishing air assault landing zones, close air support for strike aircraft "	0.1	0.5	0.1	0.5
53	"launched into geosynchronous orbit on a Titan IV booster"	0.1	0.1	0.4	*
54	"Air Force investigated Unidentified Flying Objects"	0.5	0.1	0.1	0.2
55	"vice commander, 22nd Air Force"	0.5	0.1	0.1	*
56	"the operation of the USAF Test Pilot School"	0.4	4.7	0.3	0.6
57	"1991 Defense Systems Management College"	0.1	0.1	0.6	0.4
58	"aide to the commandant Air War College"	*	0.3	0.7	*
59	"vice commander Headquarters U.S. Air Forces in Europe"	0.4	0.1	0.1	0.4
60	"25th Air Division, Tactical Air Command, McChord"	0.1	1	0.1	0.5
61	"Keeping the peace in Sarajevo"	0.1	0.1	0.3	*
62	"Pegasus launches experimental payload"	0.1	0.1	0.1	0.4
63	"Tanker units rotate in south of France"	0.1	0.4	0.1	0
64	"Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994"	*	*	*	*
65	"The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land"	0.1	0.1	0.1	0.5

66	"Top civil servants to enter education, development program"	0.4	0.1	0	0.4
67	"Air Force is a leading-edge customer"	0.1	0.4	0.1	0.5
68	"General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators"	0.1	0.1	0.1	*
69	"obtain military memorabilia such as patches, uniform items, mugs"	*	0.1	0.6	0.1
70	"Navy Seals egress from a MH-53J Pave Low in the mountains of Norway"	0.4	0.1	0.1	0.5
71	"Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2"	*	0.1	0.6	*
72	"Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)"	0.1	0.4	0.1	*
73	"Chief of Staff visits troops in Tuzla"	0.1	0.1	0.5	*
74	"The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training."	0.5	0.5	0.4	*
75	"any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations"	*	*	*	*
76	"performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world"	0.1	0.1	0.1	*
77	"It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display"	0.5	0.1	0.6	0.6
78	"The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces."	*	*	*	*

79	"General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010"	-	0.4	0.1	0.8
80	"If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices"	0.1	*	0.1	0.9

Memory Usage

Query	BNC	AF	S1	S2S4
1 SPA147	0.1	0.3	0.3	*
2 scouts	0.2	0.2	0.4	*
3 "Samuel C. Robbins"	0.1	0.3	0.2	*
4 ipms	0.3	0.2	0.1	*
5 air ADJ logistics	0.2	0.2	0.2	0.4
6 AMC AND flight AND dependent	0.1	0.2	0.2	*
7 "headquarters USAF"	0.2	-	0.3	-
8 earned	*	-	-	-
9 "Joint Staff"	*	0.3	0.4	0.4
10 "reconnaissance aircraft"	0.4	-	-	0.4
11 "early separation"	0.4	0.2	-	0.4
12 "Ellins Airbase"	0.2	0.2	0.4	*
13 "Retired Pay" AND offset	0.2	0.2	0.2	0.3
14 "Office of primary responsibility"	0.2	0.2	0.2	*
15 b36 AND peacemaker	0.3	0.3	0.2	*
16 "kenneth lowry"	0.2	0.2	0.2	*
17 fox AND edwardAND j AND 15471683	0.3	0.2	0.2	0.3
18 tuition AND assistance AND changes	*	0.4	0.3	0.3
19 reqroting	0.2	0.2	0.4	0.2
20 mechanics	0.4	0.4	0.2	0.4
21 "keflavik, iceland" AND ab	0.2	0.3	0.2	*
22 "Juan A. Marrero"	*	0.2	0.4	0.2
23 winemiller	*	0.2	0.2	0.3
24 "damage control"	-	-	*	0.4
25 VC-25A	0.2	0.2	0.4	*
26 "E-Mail Listings"	0.2	0.2	0.4	*
27 "Air Force" AND artwork AND emblem AND logo	0.3	0.3	0.4	*
28 "Michael Egan"	0.2	*	0.2	*
29 "Special Tactics Team"	0.2	-	-	*
30 "ft. hood"	0.2	0.2	0.4	0.4
31 "earnest harmon air force base"	0.3	0.3	0.2	0.3
32 stealth	*	-	-	-
33 "Space Maneuver Vehicle"	0.4	0.4	0.4	*
34 Elmendorf	0.4	0.4	0.4	0.4
35 "tandem thrust"	0.4	0.3	0.4	0.2
36 accident AND shaw	0.2	0.2	0.2	0.4
37 "Jackie Parker"	0.4	0.2	0.4	0.4
38 "military paychart"	0.2	0.2	0.3	*

39 "www.afmc.wpafb.af.mil/lean/nsn."	0.2	0.3	0.2	*
40 "10th Tactical reconnaissance Wing"	0.4	0.2	0.3	0.4
41 "General Officers"	*	0.3	0.3	0.4
42 closing AND lowry	0.2	0.2	0.2	*
43 "military retired pay center"	0.3	-	-	*
44 Lompoc	0.2	*	0.4	*
45 "afi 33-270"	0.2	0.1	0.2	0.3
46 412	0.2	0.2	0.2	0.4
47 Jackson AND Charles	0.2	0.2	0.4	0.3
48 elliott	0.4	0.3	0.2	*
49 50th AND Air AND Force AND Logo	0.4	0.2	0.2	*
50 air AND force AND instruction AND 32-7064	0.3	0.1	0.2	0.2
51 "Presidential Maintenance Branch"	0.4	0.3	0.2	0.4
52 "air traffic control for establishing air assault landing zones, close air support for strike aircraft "	0.2	0.3	0.2	0.4
53 "launched into geosynchronous orbit on a Titan IV booster"	0.2	0.2	0.4	*
54 "Air Force investigated Unidentified Flying Objects"	0.4	0.2	0.2	0.3
55 "vice commander, 22nd Air Force"	0.4	0.2	0.2	*
56 "the operation of the USAF Test Pilot School"	0.4	0.4	0.4	0.4
57 "1991 Defense Systems Management College"	0.2	0.2	0.4	0.4
58 "aide to the commandant Air War College"	*	0.3	0.4	*
59 "vice commander Headquarters U.S. Air Forces in Europe"	0.3	0.2	0.2	0.3
60 "25th Air Division, Tactical Air Command, McChord"	0.2	0.3	0.2	0.4
61 "Keeping the peace in Sarajevo"	0.2	0.1	0.4	*
62 "Pegasus launches experimental payload"	0.2	0.2	0.2	0.4
63 "Tanker units rotate in south of France"	0.2	0.3	0.2	0.2
64 "Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994"	*	*	*	*
65 "The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land"	0.2	0.2	0.2	0.4

66 "Top civil servants to enter education, development program"	0.4	0.2	0.2	0.4
67 "Air Force is a leading-edge customer"	0.1	0.3	0.2	0.4
68 "General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators"	0.2	0.2	0.1	*
69 "obtain military memorabilia such as patches, uniform items, mugs"	*	0.3	0.4	0.2
70 "Navy Seals egress from a MH-53J Pave Low in the mountains of Norway"	0.4	0.2	0.2	0.4
71 "Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2"	*	0.2	0.4	*
72 "Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)"	0.2	0.3	0.2	*
73 "Chief of Staff visits troops in Tuzla"	0.2	0.2	0.4	*
74 "The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training."	0.4	0.4	0.4	*
75 "any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations"	*	*	*	*
76 "performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world"	0.1	0.2	0.2	*
77 "It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display"	0.4	0.2	0.4	0.4
78 "The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces."	*	*	*	*

79	"General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010"	-	0.3	0.2	0.4
80	"If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices"	0.2	*	0.2	0.4

Appendix E: Stoplisted Queries

BNC Stoplisted Queries

Query #	Original Query	# of Words	Query after stoplisting	# of Words	# of Words stoplisted
1	SPA147	1	SPA147	1	0
2	scouts	1	scouts	1	0
3	Samuel C. Robbins	3	Samuel C. Robbins	3	0
4	ipms	1	ipms	1	0
5	air ADJ logistics	2	air ADJ logistics	2	0
6	AMC AND flight AND dependent	3	AMC AND flight AND dependent	3	0
7	headquarters USAF	2	headquarters USAF	2	0
8	earned	1	earned	1	0
9	Joint Staff	2	Joint	1	1
10	reconnaissance aircraft	2	reconnaissance aircraft	2	0
11	early separation	2	separation	1	1
12	Ellins Airbase	2	Ellins Airbase	2	0
13	Retired Pay AND offset	3	Retired AND offset	2	1
14	Office of primary responsibility	4	primary responsibility	2	2
15	b36 AND peacemaker	2	b36 peacemaker	2	0
16	kenneth lowry	2	kenneth lowry	2	0
17	fox AND edward AND j AND 15471683	4	fox AND edward AND j AND 15471683	4	0
18	tuition AND assistance AND changes	3	tuition AND assistance AND changes	3	0
19	reqroting	1	reqroting	1	0
20	mechanics	1	mechanics	1	0
21	keflavik, iceland AND ab	3	keflavik, iceland AND ab	3	0
22	Juan A. Marrero	3	Juan A. Marrero	3	0
23	winemiller	1	winemiller	1	0
24	damage control	2	damage	1	1
25	VC-25A	1	VC-25A	1	0
26	E-Mail Listings	2	E-Mail Listings	2	0
27	Air Force AND artwork AND emblem AND logo	5	Air Force AND artwork AND emblem AND logo	5	0
28	Michael Egan	2	Michael Egan	2	0

29	Special Tactics Team	3	Tactics Team	2	1
30	ft. hood	2	ft. hood	2	0
31	earnest harmon air force base	5	earnest harmon air force base	5	0
32	stealth	1	stealth	1	0
33	Space Maneuver Vehicle	3	Space Maneuver Vehicle	3	0
34	Elmendorf	1	Elmendorf	1	0
35	tandem thrust	2	tandem thrust	2	0
36	accident AND shaw	2	accident shaw	2	0
37	Jackie Parker	2	Jackie Parker	2	0
38	military paychart	2	military paychart	2	0
39	www.afmc.wpafb.af.mil/lean/nsn.	1	ww.afmc.wpafb.af.mil/lean/nsn.	1	0
40	10th Tactical reconnaissance Wing	4	10th Tactical reconnaissance Wing	4	0
41	General Officers	2	Officers	1	1
42	closing AND lowry	2	closing AND lowry	2	0
43	military retired pay center	4	military retired center	3	1
44	Lompoc	1	Lompoc	1	0
45	afi 33-270	2	afi 33-270	2	0
46	412	1	412	1	0
47	Jackson AND Charles	2	Jackson AND Charles	2	0
48	elliott	1	elliott	1	0
49	50th AND Air AND Force AND Logo	4	50th AND Air AND Force AND Logo	4	0
50	air AND force AND instruction AND 32-7064	4	air AND force AND instruction AND 32-7064	4	0
51	Presidential Maintenance Branch	3	Presidential Maintenance Branch	3	0
52	air traffic control for establishing air assault landing zones, close air support for strike aircraft	15	air traffic establishing air assault landing zones, close air strike aircraft	11	4
53	launched into geosynchronous orbit on a Titan IV booster	9	launched geosynchronous orbit Titan IV booster	6	3
54	Air Force investigated Unidentified Flying Objects	6	Air Force investigated Unidentified Flying Objects	6	0

55	vice commander, 22nd Air Force	5	vice commander, 22nd Air Force	5	0
56	the operation of the USAF Test Pilot School	8	operation USAF Test Pilot	4	4
57	1991 Defense Systems Management College	5	1991 Defense Systems Management College	5	0
58	aide to the commandant Air War College	7	aide commandant Air College	4	3
59	vice commander Headquarters U.S. Air Forces in Europe	8	vice commander Headquarters U.S. Air Forces Europe	7	1
60	25th Air Division, Tactical Air Command, McChord	7	25th Air Division, Tactical Air Command, McChord	7	0
61	Keeping the peace in Sarajevo	5	Keeping peace Sarajevo	3	2
62	Pegasus launches experimental payload	4	Pegasus launches experimental payload	4	0
63	Tanker units rotate in south of France	7	Tanker units rotate France	4	3
64	Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994	15	Yokota's Sports Fitness Center selected Air Force's 1994	8	7
65	The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land	16	nose landing gear EC-135 command aircraft collapsed attempting land	9	7
66	Top civil servants to enter education, development program	8	civil servants enter education, program	5	3
67	Air Force is a leading-edge customer	6	Air Force leading-edge customer	4	2
68	General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators	18	Ronald R. Fogleman announced professional reading program designed promote Air Force operators	12	6

69	obtain military memorabilia such as patches, uniform items, mugs	9	obtain military memorabilia patches, uniform items, mugs	7	2
70	Navy Seals egress from a MH-53J Pave Low in the mountains of Norway	13	Navy Seals egress MH-53J Pave Low mountains Norway	8	5
71	Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2	25	Crews aboard EC-130E Hercules aircraft 42nd Airborne Command Squadron participated search A-10 missing April 2	15	10
72	Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)	20	Combat Controllers (CCT) ground combat forces assigned Tactics Squadrons Air Force Operations Command (AFSOC)	14	6
73	Chief of Staff visits troops in Tuzla	7	Chief visits troops Tuzla	4	3
74	The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training.	19	U.S. Air Force committed safety minimizing collateral noise associated low-level flying training.	12	7
75	any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations	29	deny, exploit, corrupt destroy enemy's functions protecting Air Force assets actions exploiting military operations	14	15

76	performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world	18	performs precision aerial maneuvers demonstrating capabilities Air Force performance aircraft throughout	11	7
77	It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display	23	forward-looking infrared radar, missile radar warning receivers, chaff flare dispensers night-vision goggle compatible heads-up display	15	8
78	The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces.	16	secretary remarks commissioning 46 graduating Aggies officers America's armed forces.	10	6
79	General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010	25	Fogleman stated superiority vital achieving goals Chairman Joint Chiefs Joint Vision 2010	12	13
80	If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices	26	Air Force succeed modernization Quality initiatives, free resources revolution practices	10	16
Total		494		342	152

AF Stoplisted Queries

Query #	Original Query	# of Words	Query after stoplisting	# of Words	# of Words stoplisted
1	SPA147	1	SPA147	1	0
2	scouts	1	scouts	1	0
3	Samuel C. Robbins	3	Samuel C. Robbins	3	0
4	ipms	1	ipms	1	0
5	air ADJ logistics	2	ADJ	0	2
6	AMC AND flight AND dependent	3	AMC AND dependent	2	1
7	headquarters USAF	2	headquarters	1	1
8	earned	1	earned	1	0
9	Joint Staff	2		0	2
10	reconnaissance aircraft	2	reconnaissance	1	1
11	early separation	2	early separation	2	0
12	Ellins Airbase	2	Ellins Airbase	2	0
13	Retired Pay AND offset	3	Retired AND offset	2	1
14	Office of primary responsibility	4		0	4
15	b36 AND peacemaker	2	b36 AND peacemaker	2	0
16	kenneth lowry	2	kenneth lowry	2	0
17	fox AND edward AND j AND 15471683	4	fox AND edward AND j AND 15471683	4	0
18	tuition AND assistance AND changes	3	tuition AND	1	2
19	reqroting	1	reqroting	1	0
20	mechanics	1	mechanics	1	0
21	keflavik, iceland AND ab	3	keflavik, iceland AND ab	3	0
22	Juan A. Marrero	3	Juan A. Marrero	3	0
23	winemiller	1	winemiller	1	0
24	damage control	2	damage	1	1
25	VC-25A	1	VC-25A	1	0
26	E-Mail Listings	2	E-Mail Listings	2	0
27	Air Force AND artwork AND emblem AND logo	5	AND artwork AND emblem AND logo	3	2
28	Michael Egan	2	Michael Egan	2	0
29	Special Tactics Team	3	Tactics	1	2
30	ft. hood	2	ft. Hood	2	0

31	earnest harmon air force base	5	earnest harmon	2	3
32	stealth	1	stealth	1	0
33	Space Maneuver Vehicle	3	Maneuver	1	2
34	Elmendorf	1	Elmendorf	1	0
35	tandem thrust	2	tandem thrust	2	0
36	accident AND shaw	2	accident AND shaw	2	0
37	Jackie Parker	2	Jackie Parker	2	0
38	military paychart	2	paychart	1	1
39	www.afmc.wpafb.af.mil/lean/nsn.	1	www.afmc.wpafb.af.mil/lean/nsn.	1	0
40	10th Tactical reconnaissance Wing	4	10th Tactical reconnaissance Wing	4	0
41	General Officers	2		0	2
42	closing AND lowry	2	closing lowry	2	0
43	military retired pay center	4	retired	1	3
44	Lompoc	1	Lompoc	1	0
45	afi 33-270	2	33-270	1	1
46	412	1	412	1	0
47	Jackson AND Charles	2	Jackson Charles	2	0
48	elliott	1	elliott	1	0
49	50th AND Air AND Force AND Logo	4	50th AND Logo	2	2
50	air AND force AND instruction AND 32-7064	4	AND 32-7064	1	3
51	Presidential Maintenance Branch	3	Presidential Branch	2	1
52	air traffic control for establishing air assault landing zones, close air support for strike aircraft	15	traffic establishing assault landing zones, close strike	7	8
53	launched into geosynchronous orbit on a Titan IV booster	9	launched geosynchronous orbit Titan IV booster	6	3
54	Air Force investigated Unidentified Flying Objects	6	investigated Unidentified Objects	3	3
55	vice commander, 22nd Air Force	5	vice commander, 22nd	3	2

56	the operation of the USAF Test Pilot School	8	Pilot	1	7
57	1991 Defense Systems Management College	5	1991 College	2	3
58	aide to the commandant Air War College	7	aide commandant War College	4	3
59	vice commander Headquarters U.S. Air Forces in Europe	8	vice commander Headquarters U.S. Europe	5	3
60	25th Air Division, Tactical Air Command, McChord	7	25th Division, Tactical Command, McChord	5	2
61	Keeping the peace in Sarajevo	5	Keeping peace Sarajevo	3	2
62	Pegasus launches experimental payload	4	Pegasus launches experimental payload	4	0
63	Tanker units rotate in south of France	7	Tanker rotate south France	4	3
64	Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994	15	Yokota's Sports Fitness selected Force's best 1994	7	8
65	The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land	16	nose landing gear EC-135 command collapsed while attempting land	9	7
66	Top civil servants to enter education, development program	8	Top servants education,	3	5
67	Air Force is a leading-edge customer	6	leading-edge customer	2	4
68	General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators	18	Ronald R. Fogleman announced professional reading designed promote operators	9	9
69	obtain military memorabilia such as patches, uniform items, mugs	9	obtain memorabilia patches, uniform items, mugs	6	3

70	Navy Seals egress from a MH-53J Pave Low in the mountains of Norway	13	Navy Seals egress MH-53J Pave Low mountains Norway	8	5
71	Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2	25	Crews aboard EC-130E Hercules 42nd Airborne Command Squadron participated search A-10 missing since April 2	15	10
72	Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)	20	Controllers (CCT) Tactics Squadrons Command (AFSOC)	6	14
73	Chief of Staff visits troops in Tuzla	7	visits troops Tuzla	3	4
74	The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training.	19	U.S. committed to minimizing collateral noise associated low-level training.	9	10
75	any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations	29	deny, exploit, corrupt enemy's while protecting assets against exploiting own	10	19

76	performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world	18	performs precision aerial maneuvers demonstrating capabilities high people throughout world	10	8
77	It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display	23	forward-looking infrared radar, missile radar warning receivers, chaff flare dispensers night-vision goggle compatible heads-up display	15	8
78	The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces.	16	remarks while commissioning 46 graduating Aggies America's armed forces.	9	7
79	General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010	25	Fogleman stated superiority vital achieving goals Chairman Chiefs Vision 2010	10	15
80	If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices	26	succeed modernization Life initiatives, free revolution business practices	8	18
Total		494		264	230

S1 Stoplisted Queries

Query #	Original Query	# of Words	Query after stoplisting	# of Words	# of Words stoplisted
1	SPA147	1	SPA147	1	0
2	scouts	1	scouts	1	0
3	Samuel C. Robbins	3	Samuel C. Robbins	3	0
4	ipms	1	ipms	1	0
5	air ADJ logistics	2	ADJ	0	2
6	AMC AND flight AND dependent	3	AMC AND dependent	2	1
7	headquarters USAF	2		0	2
8	earned	1	earned	1	0
9	Joint Staff	2		0	2
10	reconnaissance aircraft	2	reconnaissance	1	1
11	early separation	2	early	1	1
12	Ellins Airbase	2	Ellins Airbase	2	0
13	Retired Pay AND offset	3	Retired AND offset	2	1
14	Office of primary responsibility	4		0	4
15	b36 AND peacemaker	2	b36 AND peacemaker	2	0
16	kenneth lowry	2	kenneth lowry	2	0
17	fox AND edward AND j AND 15471683	4	fox AND edward AND j AND 15471683	4	0
18	tuition AND assistance AND changes	3	tuition AND	1	2
19	reqroting	1	reqroting	1	0
20	mechanics	1	mechanics	1	0
21	keflavik, iceland AND ab	3	keflavik, iceland AND ab	3	0
22	Juan A. Marrero	3	Juan A. Marrero	3	0
23	winemiller	1	winemiller	1	0
24	damage control	2	damage	1	1
25	VC-25A	1	VC-25A	1	0
26	E-Mail Listings	2	E-Mail Listings	2	0
27	Air Force AND artwork AND emblem AND logo	5	artwork AND emblem AND logo	3	2
28	Michael Egan	2	Michael Egan	2	0
29	Special Tactics Team	3	Tactics	1	2
30	ft. hood	2	ft. hood	2	0

31	earnest harmon air force base	5	earnest harmon	2	3
32	stealth	1	stealth	1	0
33	Space Maneuver Vehicle	3	Maneuver	1	2
34	Elmendorf	1	Elmendorf	1	0
35	tandem thrust	2	tandem thrust	2	0
36	accident AND shaw	2	accident AND shaw	2	0
37	Jackie Parker	2	Jackie Parker	2	0
38	military paychart	2	paychart	1	1
39	www.afmc.wpafb.af.mil/lean/nsn.	1	www.afmc.wpafb.af.mil/lean/nsn.	1	0
40	10th Tactical reconnaissance Wing	4	10th Tactical reconnaissance Wing	4	0
41	General Officers	2		0	2
42	closing AND lowry	2	closing AND lowry	2	0
43	military retired pay center	4	retired	1	3
44	Lompoc	1	Lompoc	1	0
45	afi 33-270	2	33-270	1	1
46	412	1	412	1	0
47	Jackson AND Charles	2	Jackson AND Charles	2	0
48	elliott	1	elliott	1	0
49	50th AND Air AND Force AND Logo	4	50th AND Logo	2	2
50	air AND force AND instruction AND 32-7064	4	32-7064	1	3
51	Presidential Maintenance Branch	3	Presidential Branch	2	1
52	air traffic control for establishing air assault landing zones, close air support for strike aircraft	15	traffic establishing assault landing zones, close strike	7	8
53	launched into geosynchronous orbit on a Titan IV booster	9	launched into geosynchronous orbit on a Titan IV booster	9	0
54	Air Force investigated Unidentified Flying Objects	6	investigated Unidentified Objects	3	3
55	vice commander, 22nd Air Force	5	vice commander, 22nd	3	2

56	the operation of the USAF Test Pilot School	8	the the Pilot School	4	4
57	1991 Defense Systems Management College	5	1991 College	2	3
58	aide to the commandant Air War College	7	aide to the commandant War College	6	1
59	vice commander Headquarters U.S. Air Forces in Europe	8	vice commander U.S. in Europe	5	3
60	25th Air Division, Tactical Air Command, McChord	7	25th Division, Tactical Command, McChord	5	2
61	Keeping the peace in Sarajevo	5	Keeping the peace in Sarajevo	5	0
62	Pegasus launches experimental payload	4	Pegasus launches experimental payload	4	0
63	Tanker units rotate in south of France	7	Tanker rotate in south France	5	2
64	Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994	15	Yokota's Sports Fitness it was the Force's best 1994	9	6
65	The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land	16	The nose landing gear EC-135 command collapsed while attempting to land	11	5
66	Top civil servants to enter education, development program	8	Top servants to education,	4	4
67	Air Force is a leading-edge customer	6	is a leading-edge customer	4	2
68	General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators	18	Ronald R. Fogleman announced a professional reading designed to promote the operators	12	6

69	obtain military memorabilia such as patches, uniform items, mugs	9	obtain memorabilia patches, uniform items, mugs	6	3
70	Navy Seals egress from a MH-53J Pave Low in the mountains of Norway	13	Navy Seals egress from a MH-53J Pave Low in the mountains Norway	12	1
71	Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2	25	Crews aboard EC-130E Hercules from the 42nd Airborne Command have participated in the search A-10 missing since April 2	19	6
72	Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)	20	Controllers (CCT) ground to Tactics Squadrons the Command (AFSOC)	9	11
73	Chief of Staff visits troops in Tuzla	7	visits troops in Tuzla	4	3
74	The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training.	19	The U.S. is committed to to minimizing the collateral noise with low-level training.	13	6
75	any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations	29	to deny, exploit, corrupt the enemy's its while protecting assets against those exploiting its own	15	14

76	performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world	18	performs precision aerial maneuvers demonstrating the high to people throughout the world	12	6
77	It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display	23	It also have a forward-looking infrared radar, missile radar warning receivers, chaff flare dispensers night-vision goggle compatible heads-up display	19	4
78	The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces.	16	The made her remarks while commissioning 46 graduating Aggies in America's armed forces.	13	3
79	General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010	25	Fogleman stated that superiority be vital in achieving the goals the Chairman the Chiefs Vision 2010	16	9
80	If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices	26	the is to succeed in its modernization Life initiatives, we free up a revolution in business practices	17	9
Total		494		329	165

S2S4 Stoplisted Queries

Query #	Original Query	# of Words	Query after stoplisting	# of Words	# of Words stoplisted
1	SPA147	1	SPA147	1	0
2	scouts	1	scouts	1	0
3	Samuel C. Robbins	3	Samuel C. Robbins	3	0
4	ipms	1	ipms	1	0
5	air ADJ logistics	2	air ADJ logistics	2	0
6	AMC AND flight AND dependent	3	AND flight AND dependent	2	1
7	headquarters USAF	2	headquarters	1	1
8	earned	1	earned	1	0
9	Joint Staff	2	Joint Staff	2	0
10	reconnaissance aircraft	2	reconnaissance aircraft	2	0
11	early separation	2	early separation	2	0
12	Ellins Airbase	2	Ellins Airbase	2	0
13	Retired Pay AND offset	3	Retired Pay AND offset	3	0
14	Office of primary responsibility	4	Office of primary responsibility	4	0
15	b36 AND peacemaker	2	b36 AND peacemaker	2	0
16	kenneth lowry	2	kenneth lowry	2	0
17	fox AND edward AND j AND 15471683	4	fox AND edward AND j AND 15471683	4	0
18	tuition AND assistance AND changes	3	tuition AND assistance AND changes	3	0
19	reqroting	1	reqroting	1	0
20	mechanics	1	mechanics	1	0
21	keflavik, iceland AND ab	3	keflavik, iceland AND ab	3	0
22	Juan A. Marrero	3	Juan A. Marrero	3	0
23	winemiller	1	winemiller	1	0
24	damage control	2	damage control	2	0
25	VC-25A	1	VC-25A	1	0
26	E-Mail Listings	2	E-Mail Listings	2	0
27	Air Force AND artwork AND emblem AND logo	5	Air Force AND artwork AND emblem AND logo	5	0
28	Michael Egan	2	Michael Egan	2	0
29	Special Tactics Team	3	Special Tactics Team	3	0
30	ft. hood	2	ft. hood	2	0

31	earnest harmon air force base	5	earnest harmon air force base	5	0
32	stealth	1	stealth	1	0
33	Space Maneuver Vehicle	3	Space Maneuver Vehicle	3	0
34	Elmendorf	1	Elmendorf	1	0
35	tandem thrust	2	tandem thrust	2	0
36	accident AND shaw	2	accident AND shaw	2	0
37	Jackie Parker	2	Jackie Parker	2	0
38	military paychart	2	military paychart	2	0
39	www.afmc.wpafb.af.mil/lean/nsn.	1	www.afmc.wpafb.af.mil/lean/nsn.	1	0
40	10th Tactical reconnaissance Wing	4	10th Tactical reconnaissance Wing	4	0
41	General Officers	2	General Officers	2	0
42	closing AND lowry	2	closing AND lowry	2	0
43	military retired pay center	4	military retired pay center	4	0
44	Lompoc	1	Lompoc	1	0
45	afi 33-270	2	33-270	1	1
46	412	1	412	1	0
47	Jackson AND Charles	2	Jackson AND Charles	2	0
48	elliott	1	elliott	1	0
49	50th AND Air AND Force AND Logo	4	50th AND Air AND Force AND Logo	4	0
50	air AND force AND instruction AND 32-7064	4	air AND force AND instruction AND 32-7064	4	0
51	Presidential Maintenance Branch	3	Presidential Maintenance Branch	3	0
52	air traffic control for establishing air assault landing zones, close air support for strike aircraft	15	air traffic control for establishing air assault landing zones, close air support for strike aircraft	15	0
53	launched into geosynchronous orbit on a Titan IV booster	9	launched into geosynchronous orbit on a Titan IV booster	9	0
54	Air Force investigated Unidentified Flying Objects	6	Air Force investigated Unidentified Flying Objects	6	0

55	vice commander, 22nd Air Force	5	vice commander, 22nd Air Force	5	0
56	the operation of the USAF Test Pilot School	8	the operation of the Test Pilot School	7	1
57	1991 Defense Systems Management College	5	1991 Systems Management College	4	1
58	aide to the commandant Air War College	7	aide to the commandant Air War College	7	0
59	vice commander Headquarters U.S. Air Forces in Europe	8	vice commander Headquarters U.S. Air Forces in Europe	8	0
60	25th Air Division, Tactical Air Command, McChord	7	25th Air Division, Tactical Air Command, McChord	7	0
61	Keeping the peace in Sarajevo	5	Keeping the peace in Sarajevo	5	0
62	Pegasus launches experimental payload	4	Pegasus launches experimental payload	4	0
63	Tanker units rotate in south of France	7	Tanker units rotate in south of France	7	0
64	Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994	15	Yokota's Sports and Fitness Center when it was selected the Air Force's best for 1994	15	0
65	The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land	16	The nose landing gear of an EC-135 command and control aircraft collapsed while attempting to land	16	0
66	Top civil servants to enter education, development program	8	Top civil servants to enter education, development program	8	0
67	Air Force is a leading-edge customer	6	Air Force is a leading-edge customer	6	0
68	General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators	18	General Ronald R. Fogleman announced a professional reading program designed to promote the development of Air Force operators	18	0

69	obtain military memorabilia such as patches, uniform items, mugs	9	obtain military memorabilia such as patches, uniform items, mugs	9	0
70	Navy Seals egress from a MH-53J Pave Low in the mountains of Norway	13	Navy Seals egress from a MH-53J Pave Low in the mountains of Norway	13	0
71	Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2	25	Crews aboard EC-130E Hercules aircraft from the 42nd Airborne Command and Control Squadron have participated in the search for an A-10 missing since April 2	25	0
72	Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)	20	Combat Controllers (CCT) are ground combat forces assigned to Special Tactics Squadrons within the Air Force Special Operations Command (AFSOC)	20	0
73	Chief of Staff visits troops in Tuzla	7	Chief of Staff visits troops in Tuzla	7	0
74	The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training.	19	The U.S. Air Force is committed to safety and to minimizing the collateral noise associated with low-level flying training.	19	0
75	any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations	29	any action to deny, exploit, corrupt or destroy the enemy's information and its functions while protecting Air Force assets against those actions and exploiting its own military information operations	29	0

76	performs precision aerial maneuvers demonstrating the capabilities of Air Force high performance aircraft to people throughout the world	18	performs precision aerial demonstrating the capabilities of Air Force high performance aircraft to people throughout the world	17	1
77	It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display	23	It will also have a forward-looking infrared radar, missile and radar warning receivers, chaff and flare dispensers and night-vision goggle compatible heads-up display	23	0
78	The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces.	16	The secretary made her remarks while commissioning 46 graduating Aggies as officers in America's armed forces.	16	0
79	General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010	25	General Fogleman stated that information superiority will be vital in achieving the goals of the Chairman of the Joint Chiefs of Staff Joint Vision 2010	25	0
80	If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices	26	If the Air Force is to succeed in its modernization and Quality of Life initiatives, we must free up resources through a revolution in business practices	26	0
Total		494		488	6

Bibliography

"Adam Kilgarriff Reasearch Fellow", n. pag WWWeb http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/. 18 Jan 1996.

Edmundson, H. P. and R. E. Wyllys. Automatic Indexing and Abstracting of the Contents of Documents. Contract RADC-TR-59-20S. Planning and Research Corporation. 31 October 1959.

Fox, C. "A Stoplist for General Text" SIGIR Forum 24, 19-35, (Fall 89/Winter 90).

Francis, W. N. and Henry Kucera. Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifilin Company, 1982.

Luhn, H. P. "The Automatic Creation of Literature Abstracts", IBM Journal, 159-163 (April 1958)

Maxwell, Joseph A. "Qualitative Research Design: An Interactive Approach," Applied Social Research Methods Series, 41: 86-98 (1996).

Sinclair, Hohn. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.

Snoddy, David W. Records Analysis and Classification System: A Proof of Concept System for the Automated Classification of United States Air Force Records. MS Thesis, AFIT/GIR/LAR/96D-11. School of Logistics and Acquisition Management , Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1996 (AD-A319714).

"The LEXIS-NEXIS Computing Complex.", n. pag. WWWeb, http://www.lexisnexis.com/lncc/about/datacenter.html. 8 October 1997.

van Rijsbergen, C. J. Information Retrieval (Second Edition). London: Butterworth, 1979.

Vita

Captain Craig N. Berg was born on 13 May 1964 in Seaside, Oregon. He graduated from Taft High School, Lincoln City, Oregon, in 1982. He graduated from the University of Portland, Portland, Oregon, in 1987 with a Bachelor of Science in Electrical Engineering. Upon graduation he was commissioned through the Air Force Reserve Officer Training Corps. In October of 1987 he began the Basic Communications Officer Training course, Keesler Air Force Base, Mississippi. In February 1988 he was assigned to Headquarters Strategic Air Command as a Plans and Exercise Staff Officer, in the Contingency Communications Branch. In January of 1990 he became the Chief of Special Operations Communications. On 13 April 1991 he was reassigned to Headquarters United States Forces Japan, Yokota Air Base, Japan. In August of 1993 he moved to the 374th Communications Squadron at Yokota Air Base as the Operations Flight Commander.

In May 1996, Captain Berg entered the School of Logistics and Acquisition Management, Air Force Institute of Technology.

Captain Berg is married to the former Susan M. Sullivan of Vancouver, Washington.

Permanent Address: 608 Ash
Vancouver, WA, 98661

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per reponse, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE December 1997	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE
DEVELOPING A CORPUS SPECIFIC STOPLIST USING QUANTITATIVE COMPARISON

5. FUNDING NUMBERS

6. AUTHOR(S)
Craig N. Berg, Captain, USAF

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
2950 P Street
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GIR/LAL/97D-2

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Technical Information Center
Research, Development & Acquisition
Information Support Directorate (DTIC-A)
8725 John J. Kingman Road
Ft Belvoir VA, 22060-6218

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

We have become overwhelmed with electronic information and it seems our situation is not going to improve. It is becoming increasingly common for people to work with information on a daily basis. We seem to spend more and more time looking for information, and it is taking longer because more information is available. This thesis will look at how we can provide faster access to the information we want to find. Today's requirements are closely related to searching for information using queries. At the heart of the query process is the removal of search terms having little or no significance to the search being performed. Words considered to have little significance, in terms of their searching power, called stopwords, are compiled in a stoplist. Stoplists are usually constructed from commonly occurring words in the English language. This approach is acceptable for systems handling broad categories of information. We will build a stoplist for a specific area of interest based on a specific body of linguistic data, or corpus. A stoplist developed from an Air Force corpus will be tested to see if it is more effective than a stoplist created from a general use corpus.

14. Subject Terms
Text Processing, Data Bases, Internet, Abstracts

15. NUMBER OF PAGES
136

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

AFIT Control Number AFIT/GIR/LAL/97D-2

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. **Please return completed questionnaire** to: AIR FORCE INSTITUTE OF TECHNOLOGY/LAC, 2950 P STREET, WRIGHT-PATTERSON AFB OH 45433-7765. Your response is **important.** Thank you.

1. Did this research contribute to a current research project? a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?
a. Yes b. No

3. **Please estimate** what this research would have cost in terms of manpower and dollars if it had been accomplished under contract or if it had been done in-house.

Man Years____________ $____________

4. Whether or not you were able to establish an equivalent value for this research (in Question 3), what is your estimate of its significance?

a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments (Please feel free to use a separate sheet for more detailed answers and include it with this form):

____
Name and Grade

____
Organization

____
Position or Title

____
Address